mathematics

MDPI

Article

Polling Systems and Their Application to Telecommunication Networks

Vladimir Vishnevsky *,† and **Olga Semenova** †

Institute of Control Sciences of Russian Academy of Sciences, 117997 Moscow, Russia; olgasmnv@gmail.com
* Correspondence: vishn@inbox.ru
† The authors contributed equally to this work.

Abstract: The paper presents a review of papers on stochastic polling systems published in 2007–2020. Due to the applicability of stochastic polling models, the researchers face new and more complicated polling models. Stochastic polling models are effectively used for performance evaluation, design and optimization of telecommunication systems and networks, transport systems and road management systems, traffic, production systems and inventory management systems. In the review, we separately discuss the results for two-queue systems as a special case of polling systems. Then we discuss new and already known methods for polling system analysis including the mean value analysis and its application to systems with heavy load to approximate the performance characteristics. We also present the results concerning the specifics in polling models: a polling order, service disciplines, methods to queue or to group arriving customers, and a feedback in polling systems. The new direction in the polling system models is an investigation of how the customer service order within a queue affects the performance characteristics. The results on polling systems with correlated arrivals (MAP, BMAP, and the group Poisson arrivals simultaneously to all queues) are also considered. We briefly discuss the results on multi-server, non-discrete polling systems and application of polling models in various fields.

Keywords: polling systems; polling order; queue service discipline; mean value analysis; probability-generating function method; broadband wireless network

check for updates

Citation: Vishnevsky, V.; Semenova, O. Polling Systems and Their Application to Telecommunication Networks. *Mathematics* **2021**, *9*, 117. https://doi.org/10.3390/math9020117

Received: 14 December 2020
Accepted: 4 January 2021
Published: 7 January 2021

1. Introduction

Due to the applicability of stochastic polling models, researchers face new and more complicated polling models. Stochastic polling models are effectively used for performance evaluation, design and optimization of telecommunication systems and networks, transport systems, and road management systems, traffic, production systems and inventory management systems, etc. (see [1–5]).

Polling systems are a special kind of queuing systems with multiple queues and a server (probably multiple servers common to all queues). Each queue has its input of customers. The server visits the queues in a specific manner and serves its customers. The rule to select the next queue for a visit is called *a service order*. It can be, e.g., cyclic polling when the server visits queues in a cyclic manner starting from the first queue to the last one and then returns to the first queue, or a random polling order where the next queue to be served is randomly selected.

The queues in a polling system are served accordingly to *a service discipline*, determining the number of customers that can be served during a server's visit to the queue. The most popular service disciplines are *the exhaustive service* when customers are served until the queue becomes empty, *the gated service* when the server serves only customers presented in the queue at the polling moment, with the other customers arriving during the queue service period being served at the next visit, and *the limited service* where the number of customers to be served during the server's visit is limited by a fixed or a random number.

Theoretical results obtained in the field of the polling system analysis before 1985 were described in detail in a book by Takagi [6]. The further development of theoretical results in this area published before 1995 is presented in by Borst [7], and the papers published in 1996–2009 are systematized in a review by Visnevsky and Semenova [8] and its update [9]. The review [8] is continued in the book by Vishnevsky and Semenova [1], where we emphasize how polling systems can be applied for design of broadband wireless networks and present new models describing the functioning of broadband wireless Wi-Fi and Wi-MAX networks with PCF (Point Coordination Function).

The aim of the present paper is an overview of results published in 2007 through 2020. We describe the directions of theoretical research development in the area, systematize the recent results obtained and point out new practical applications of polling systems. We also note some unsolved theoretical problems and propose to apply the machine-learning method to solve them. The paper continues reviews [8,9] and a paper by Borst and Boxma [10] which provides an overview of key research methods for cyclic polling systems with a single server, sets forth new approximate methods for systems in heavy load conditions and systems with many queues and discusses several complex unsolved problems in the field of analysis of polling systems.

The structure of the review is as follows. Section 2 gives a classification of polling systems and describes the primary polling model. Section 3 presents the results of the investigation of two-queue systems as a particular case of polling models. Section 4 further describes new methods and the results obtained by the known methods. In the framework of this review, we emphasize the mean value analysis and briefly describe its application to systems in heavy load conditions to obtain the approximated performance characteristics. The existence of a stationary mode for new polling systems is discussed in Section 5. Next, we group the publications accordingly to various aspects of polling models, in particular, a polling order (Section 6), a service discipline (Section 7), the methods to queue arriving customers (Section 8), and customer feedback (Section 9). In Section 10, we primarily focus on the papers investigating how the performance characteristics depend on the customer service order within a queue (Section 10). Systems with correlated arrivals are discussed in Section 11. The results for polling systems with multiple queues and systems in heavy load are briefly discussed in Sections 12 and 13. Section 14 observes the results concerning continuous polling models (systems where arriving customers are placed on a circle, systems with a denumerable number of queuing places, and fluid models). Please note that this is a specific and intricate field of research presented by just a few papers. Conclusions are made in Section 15.

2. Classification of Polling Systems

In this section, we describe the classification of polling systems [8] according to the number of queues, a polling order, and a queue service discipline.

Polling systems can be *discrete* (the number of queues is finite or denumerable) and *continuous* (the number of queues or the total number of waiting places in the system is nondenumerable) considering the number of queues in the system. In the latter case, systems are usually presented by a circle or a two-dimensional region where arriving customers are placed on. We also refer to fluid polling models where the amount of work is increased continuously as a fluid level at each queue. The amount of work in a queue usually means the time the server will spend in the queue while serving the queue load.

Discrete polling systems are characterized by:

- the number of queues and servers (single server or multi-server systems),
- queue parameters (arrival and service processes, service order within a queue),
- parameters of the server switchover between queues,
- service order,
- queue service disciplines
- and, probably, other parameters or the system topology.

Let the queues be numbered from 1 to N, where N is the number of queues in the system, ($N \geq 2$). The queue number i is denoted by Q_i, $i = \overline{1, N}$.

Let us give more details on a polling order and the service disciplines. *Polling order* is a rule for the server to select the next queue to visit. The *visit* to a queue means the polling of the queue and its further service (if the queue is not empty). Before visiting the queue, the server usually needs time to leave the previous queue and to prepare for service (to switch, move, etc.). This time is called the *server switchover time* between queues. *Polling* of a queue means the moment when the server has finished switching and is ready to start service in the queue. Polling a queue can have different senses depending on the system specifics (for example, the information on the queue length becomes available only at the polling moments). In the most polling models, polling times are convenient to be the embedded points of time to construct the embedded Markov processes. If there are customers in the queue, the server serves them accordingly to the service discipline prescribed to the queue (the disciplines will be discussed later in detail) in a customer service order (for example, First Come–First Served, Last Come–Last Served, random order, etc.). After serving the queue, the server leaves it and switches to the next queue.

A polling order can be:

1. *Cyclic*: the server polls queues in a cyclic manner from Q_1 to Q_N and then gets back to poll Q_1. The time the server polls the queues from Q_1 to Q_N is called a *cycle*.
2. *Periodic polling* is defined by the polling table $T = (T(1), T(2), \ldots, T(M))$ of size M ($M \geq N$), $T(i) \in \{1, \ldots, N\}, i = \overline{1, M}$. The server visits the queues in order $Q_{T(1)}$, $Q_{T(2)}$, $\ldots$, $Q_{T(M)}$, $Q_{T(1)}$, $Q_{T(2)}$, $\ldots$, $Q_{T(M)}$, $\ldots$. It is assumed that the polling table contains all queue numbers.

   The particular case of the periodic polling is a *star polling* where one queue is a priority (say Q_1), and the server visits it each time after serving another queue, namely in order Q_1, Q_2, Q_1, Q_3, $\ldots$, Q_1, Q_N. Here we also note an *elevator polling* where the queues are visited from the first to the last (up cycle), and then starting from the last queue, the server visits queues backwards (down cycle).
3. *Random polling*: the next queue to visit (say Q_i) is chosen randomly with probability p_i, $i = \overline{1, N}$, $\sum_{i=1}^{N} p_i = 1$. The random polling can be Markovian when the next queue to visit is Q_j given that the server leaves queue Q_i with probability p_{ij}, $i, j = \overline{1, N}$, $\sum_{j=1}^{N} p_{ij} = 1$, $i = \overline{1, N}$.
4. *Cyclic adaptive polling*: a server polls the queues but skips those of them, which were empty at their polling moment at the previous cycle.
5. *Priority polling*, when the queues are of different priority levels, and a queue can be served only when all higher-priority queues are empty.

Please note that the adaptive cyclic and priority polling order depend on a current system state, and the server decides which queues to visit at the specific time moments depending on full or partial information about the system state.

The queue service discipline is the number of customers which can be served during the server visit to the queue:

1. *Exhaustive service*, the server serves a queue until it becomes empty.
2. *Gated service*, the server serves only those customers in the queue which arrived before the polling moments. It can be presented as a gate in the queue, and all customers arriving before the polling moment are placed before the gate. At the polling moment, the gate is closed, and customers arriving during the queue service time are blocked (placed behind a gate) and will be served during the next visit. The special case of the gated service is a *globally gated service*, where all gates are closed simultaneously at the cycle beginning (usually it is the first queue polling moment).

   In case an arriving customer must wait in the queue several cycles before it gets served, the gated service discipline is called *multi-phase gated service*. Such discipline can be interpreted as follows. The queue has k buffers (k gates), and an arriving customer is placed at buffer k (before gate k, $k \geq 1$). At the moment the queue is polled, all customers from buffer i are moved to buffer $i-1$ ahead. During the visit time, the server serves customers only from the first buffer. Thus, a customer consecutively moves through buffers during k cycle until it is served.

3. *Limited service*: the number of customers that can be served is limited by l (l-limited service), or the time the server can spend at the queue is limited (T-limited service). Both disciplines can be exhaustive or gated. With the exhaustive case, the server leaves a queue if the queue is empty or l customers are served (the time to visit the queue has expired for T-limited service), whichever occurs first. With limited gated service, the sentence «queue is empty» is replaced by «all customers before the gate are served». A special case of $l = 1$ is sometimes called *non-exhaustive* service. T-limited service needs to specify the server behavior when the time to spend at the queue is expired: the server leaves the queue immediately, and a customer on the server leaves the system unserved or waits for re-service when the server visits the queue the next time.

   The value l, limiting the number of customers that can be served per visit, can be fixed or random. The latter case is called the randomly limited service and value l is defined at every polling moment as a value of a discrete random variable with distribution $\{a_j, j \geq 1\}$, i.e., $P\{l = j\} = a_j$. In the case of T-limited service, variable T can have an arbitrary distribution function. An example of a randomly limited service is a binomial service with l having a binomial distribution with parameters X and p. X is the queue length at the polling moment, p is a parameter, $0 < p \leq 1$. For this discipline

   $$a_j = \frac{X!}{j!(X-j)!} p^j (1-p)^{X-j}, \quad j = \overline{1, X},$$

   $a_j = 0$ for$j > X$. In practice, this discipline is described as follows. At the polling moment, each customer in the queue is marked with probability p is marked and will be served during the visit time, and with the additional probability $1-p$, it is marked to stay at the queue until the next marking procedure. This discipline can be exhaustive or gated. For the binomially exhaustive service, each customer arriving at the queue during the server visit is marked to be served during the current visit time or to stay at the queue until the next marking procedure at the next polling moment. Another example of random service is the *Bernoulli* discipline, where the first customer in the queue is served with probability 1, and then the server takes the next customer with probability p and leaves the queue with probability $1-p$. For this discipline we have $a_j = (1-p)p^{j-1}, j \geq 1$.
4. *l-decrementing service* where the server serves the queue in the queue until its length is l less than it was at the polling moment or until the queue is empty whichever occurs first, $l \geq 1$. In case $l = 1$ the discipline is called *semi-exhaustive*.
5. *Threshold service*, where the server starts serving a queue only if its length is no less than the specified level (threshold).

If all queues of a polling system have different service disciplines, the polling system is said to have a mixed service.

The object of study of the most papers on polling systems is the system with a single server and N ($N \geq 2$) queues with an unlimited waiting space. Queue Q_i has the Poisson arrival of customers with parameter λ_i. The service time of customers in Q_i are i.i.d. with distribution function $B_i(t)$, the mean $b_i = \int\limits_0^\infty t dB_i(t)$, the second moment $b_i^{(2)}$ and the Laplace Stieltjes transform (LST) $\tilde{B}_i(x) = \int_{i=0}^{\infty} e^{-xt} dB_i(t)$, $i = \overline{1, N}$. The arrival and service processes are supposed to be independent. Following to the Kendall's classification, such system is called the $M/G/1$-type polling system.

The server visits the queues at a specific polling order and serves them following to their service disciplines. The time to switch to queue Q_i (prepare for service, etc.) has a distribution function $S_i(t)$, the mean s_i, the second moment $s_i^{(2)}$ and LST $\tilde{S}_i(x)$, $i = \overline{1, N}$.

For systems with cyclic or periodic polling of queues, we also denote by s and $s^{(2)}$ the first and second moments of the total duration of server switching in one cycle:

$$s = \sum_{j=1}^{N} s_j, \quad s^{(2)} = s^2 + \sum_{j=1}^{N} (s_j^{(2)} - s_j^2).$$

Denote by $\rho_i = \lambda_i b_i$ the load of queue Q_i, by $\rho = \sum_{i=1}^{N} \rho_i$ is the total system load. The critical parameter of the cyclic polling system is cycle time. For cyclic (periodic) polling systems, the cycle is the time required for the server to visit queues from Q_1 to Q_N (from $Q_{T(1)}$ to $Q_{T(M)}$).

For cyclic polling systems, the mean cycle consists of time when the server serves the queues (it spends a fraction of the time ρ) and the time when it switches between queues (it spends the average time s per cycle). Thus, $C = \rho C + s$ which yields

$$C = \frac{s}{1-\rho}.$$

3. Two-Queue Polling Systems

In this section, we briefly outline the results of two-queue system analysis as a particular case of polling systems. Such models are usually considered in the case when the corresponding systems with an arbitrary number of queues are not provided with an accurate analysis yet, for example, in case of correlated input or when a polling model (with two or three queues) has a precise, practical application.

Winands et al. [11] consider the system with two $M/G/1$-type queues (priority and non-priority). The priority queue is served exhaustively; the non-priority queue has k-limited service. The server takes a switchover time to a queue only if the queue is not empty. It is shown that in the case of the limited-service discipline, it is possible to reduce the cost of the system operation significantly. The system is analyzed by using the probability-generating function (PGF) method, and the stationary state probabilities are found as a solution of the linear equations, which allows obtaining the customer sojourn time.

Boon et al. [12] consider a system of two queues, one of which received two priority inputs of customers. The discipline of service is exhaustive, gated, or globally gated. For this model, the authors obtain the cycle time distribution and the queue length distribution at the polling moments and provide the waiting time analysis. Vlasiou et al. [13] assume that the server switchover time to queues and the customer service times are correlated and describe two ways to correlate the switchover times. For the first way, the switchover times are defined by the sojourn times of a Markov chain in its states, and the second way is to use the two-dimensional Laplace distribution. For more information on priority polling systems, see Section 6.4.

Chernova et al. [14] investigate an $M/G/1$-type polling system where the first queue service discipline depends on the second queue state. If the second queue is not empty at the polling moment, the server serves it exhaustively, and the first queue receives a 1-limited service (the server serves no more than one customer per visit at a queue). In this case, the polling cycle is called regular. Otherwise, the first queue receives k-limited service, and the cycle is called modified. The system is investigated by the PGF method, a stability criterion, and the mean cycle time is obtained.

Dorsman et al. [15] apply the PGF method to a two-queue system with exhaustive service and a random Markovian polling order and obtain the LST of the waiting times distribution.

Boon and Winands [16] consider a two-queue $M/M/1$ type system with k-limited service and zero switchover times. They show that as a load of any queue increases (say, ρ_1 for Q_1) and the system becomes critically loaded ($\rho = \rho_1 + \rho_2 \to 1$) the other queue behaves as the corresponding $M/M/1$-type queuing system with server vacations having the k_2-Erlang distribution. It is also shown that the behavior of a queue is independent of the other's one.

The system of two $M/M/1$ non-symmetric queues is considered by Adan et al. [17]. An arriving customer joins the shortest queue. The stochastic process describing the system behavior is analyzed by using two approaches: the compensation approach developed by I. Adan for the nearest-neighbor two-dimensional random walks and the PGF method with a further reduction to a boundary value problem.

The papers by Gaidamaka and Shorgin et al. [18,19] consider the system of two $M/M/1/R$-type queues. If the number of customers in the first queue reaches a critical level of load, then the arrival parameter of the second queue is reduced. A numerical method is proposed for calculating the stationary probability distribution of the system states. Avrachenkov et al. [20] consider the following service policy for an $M/M/1/K$-type system. One queue, say Q_2, is assigned a threshold, and when the server visits Q_1 and the queue Q_2 length exceeds its threshold, the server stops serving Q_1 and switches to Q_2. The queue is served until queue length becomes lower than the threshold level, and the server gets back to queue Q_1. This study is continued by Perel and Yechiali [21] for the case $k = \infty$.

A system with a similar threshold strategy for switching servers between queues is discussed by Jolles et al. [22]. A threshold is assigned to each queue, and as the queue is served and its length becomes lower than the threshold, the server decides to continue serving the queue or leave it and switch to another queue. The authors use a matrix-analytical approach (see Section 11 to obtain the main performance characteristics). Furthermore, Perel et al. [23] consider a system where a server selects the longest queue to serve, apply the PGF method, and provide a comparative analysis to a corresponding $M/G/1$-type polling system.

A two-queuing polling system inventory model is presented by Granville and Drekic [5]. In the model, a queue is represented not by customers but by breakdowns of a limited number of identical machines that are eliminated by a general mechanic.

Here, we also note the papers on the analysis of tree-queue systems. Chernova et al. [24] establish the stability condition for the three-queue system with limited service. Liu et al. [25] consider the $M/M/1$-type system. Queue Q_2 has a higher priority than Q_3, and Q_1 is of the highest priority. Q_1 is served exhaustively, and the server interrupts serving any other queue each time a customer arrives to Q_1. Queue Q_2 is also assigned a threshold, and if the queue length exceeds the threshold at the moment the server visits Q_3, the server leaves Q_3 and switches to Q_2. The customer service at Q_3 resumes only when Q_2 length becomes less its threshold, and queue Q_1 is empty.

4. Methods to Study Polling Systems

4.1. Mean Value Analysis

The mean value analysis is a new method to study polling systems proposed by Winands et al. [26] to calculate the mean queue lengths at an arbitrary time. The mean can be applied for systems for the mean queue visit period can be calculated. Below we briefly describe the method. The duration of the server's visit to a queue is the sum of the service time of customers in the queue and the previous (for exhaustive) and the following (for gated service) server switchover time. Since the purpose of the method is to calculate the performance characteristics at an arbitrary time, it is necessary to know the expected duration of the queue visit by the server and the expected extra visit time.

The mean residual and passed service times of a customer in queue Q_i are equal and defined as $R_{B_i} = \frac{b_i^{(2)}}{2b_i}$, and the mean residual switchover time to Q_i is $R_{S_i} = \frac{s_i^{(2)}}{2s_i}$.

Let v_i be the mean visit time of the server to Q_i and the (i,j) period be the sum of j consecutive visit times of the server to queues starting from Q_i: $v_{i,j} = \sum_{n=i}^{i+j-1} v_n, \quad i,j = \overline{1,N}$. The mean fraction $q_{i,j}$ of time the system spends in state (i,j) is defined as $q_{i,j} = \frac{v_{i,j}}{C}$. The mean residual duration of (i,j) period is $R_{V_{i,j}} = \frac{v_{i,j}^{(2)}}{2v_{i,j}}$ and is unknown.

The unknown values $R_{V_{i,j}}, i,j=\overline{1,N}$ are related to $L_{i,j}, i,j=\overline{1,N}$ (the mean length of Q_i at an arbitrary moment of Q_j visit time). For a system with exhaustive service these relations have the form

$$\sum_{n=1}^{N} q_{n,1} L_{i,n} = \frac{\lambda_i}{1-\rho_i}\Big(\rho_i R_{B_i} + \frac{s_i}{C} R_{S_i} + (1-q_{i,1})(R_{V_{i+1,N-1}} + s_i)\Big). \qquad (1)$$

The relation (1) forms the system of N^2 linear equations for $2N^2$ unknowns $L_{i,j}$ and $R_{V_{i,j}}$. The rest N^2 equations are obtained by analyzing the average residual time of the period (i,j) and have the form

$$R_{V_{i,j}} = \frac{q_{i,1}}{q_{i,j}}\left(\frac{R_{V_{i,1}}}{\prod_{n=1}^{j-1}(1-\rho_{i+n})} + \sum_{n=1}^{j-1}\frac{s_{i+n}+L_{i+n,i}b_{i+n}}{\prod_{m=n}^{j-1}(1-\rho_{i+m})}\right) + \left(1-\frac{q_{i,1}}{q_{i,j}}\right)R_{V_{i+1,j-1}}. \qquad (2)$$

The derivation of (2) is described detail in [26]. The case of gated service requires the derivation of $2N(N+1)$ equations since during the visit period the queue length splits up two queues (before and behind the gate): the first is the number $\bar{L}_{i,i}$ of customers that will be served during the current server visit, and the second one is the number $\tilde{L}_{i,i}$ of customers arriving during the server's visit to the queue and they should wait for the next cycle to be served at the next server's visit.

The mean value analysis can be extended to the following polling systems: systems with a group Poisson input, systems with periodic polling order, discrete-time systems, and it also can be applied for comparative analysis of various polling models, see, e.g., [27–29].

Van Vuuren and Winands [27] apply the mean value analysis to approximate the mean waiting times for the limited service. The main idea of approximations is to split the initial system with N queues into N single-queue systems with server vacations and k-limited-service discipline. In addition, since it is most likely that a long (short) service period will be followed by a long (short) intervisit period, it is assumed that the duration of the intervisit periods is correlated with the number of customers served during the previous queue visit. The main aim of the analysis is to find the first two moments of the conditional intervisit period provided that l customers were served in the queue at the previous service period, as well as to find the intervisit period distribution.

Wierman et al. [30] apply the mean value analysis to the system with exhaustive or gated service under a variety of the customer service order within a queue: FCFS (First Come–First Served, as a particular case considered by Winands et al. [26]), LCFS (Last Come–First Served) and preemptive LCFS, Processor Sharing, Shortest Job First and Shortest Remaining Processing Time. For an $M/G/1$-type polling system of with variable customer arrival parameters depending on the server position, Boon et al. [31] consider the mean value analysis to obtain the mean waiting times in queues and residual average cycle time, as well as the LST of the joint number of customers in the system at polling moments, the waiting time the cycle time distributions.

The mean value analysis is applied by Vishnevsky and Semenova [29] for a system with adaptive dynamic polling. Adaptive polling assumes that the server skips (does not poll) the queues that were empty at their polling moments in the previous cycle. If all queues of the system should be skipped, the server takes a vacation and then starts polling all queues in the cycle. The analysis is based on an approximate calculation of the probabilities that the queue will be skipped in a cycle followed by the application of the mean value analysis to calculate the mean waiting times.

Vishnevsky et al. [32] propose a duplex polling system describing the performance of the polling systems in high-speed wireless mesh networks. The queues are polled cyclically by two servers. Some of the queues are available for both servers, and the remaining queues are polled just by one of the servers.

4.2. Generating Function Method

The probability-generating function (PGF) method to study cyclic polling systems with exhaustive, gated and globally gated service is described in detail by Yechiali [33]. The method is widely used for the analysis of various polling systems, and in this section we list a few papers using this method. The remaining papers are discussed in the other sections of this review taking into account the specifics of their polling models. For a cyclic $M/G/1$polling system, Boxma et al. [34] remove the restrictions on the type of service and obtain the relations for the PGF of the system states at embedded moments (at the polling moments, the moments the server starts service and leaves a queue, the moments of a customer service beginning and completion). Switchover times can be zero and non-zero. In case of a zero switchover time, it is assumed that when the system becomes empty, the server stops at the first queue and starts polling the queues again as a new customer arrives to the system. The main result of this analysis is an expression for the generating function $Q(\mathbf{z}) = Q(z_1, z_2, ..., z_N)$ of the joint queue length distribution at an arbitrary time:

$$Q(\mathbf{z}) = \frac{1}{C}\sum_{i=1}^{N}\left(\frac{V_{b_i}(\mathbf{z}) - V_{c_i}(\mathbf{z})}{\sum_{j=1}^{N}\lambda_j(1-z_j)}\,\frac{z_i\left(1-\tilde{B}_i\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right)\right)}{z_i - \tilde{B}_i\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right)} + \frac{V_{c_i}(\mathbf{z}) - V_{b_{i+1}}(\mathbf{z})}{\sum_{j=1}^{N}\lambda_j(1-z_j)}\right)$$

where the mean cycle time C for zero switchover time is $C = s/(1-\rho)$, and $C = \frac{V_{b_1}(\mathbf{0})}{\lambda(1-\rho)}$ otherwise, $V_{b_i}(\mathbf{z})$ is the PGF of the joint queue length at Q_i polling moments, $V_{c_i}(\mathbf{z})$ is the PGF of the joint queue length at the moments the server departs from Q_i.

The LST of the total amount of work in the system at an arbitrary time is derived from

$$\tilde{X}(\omega) = \frac{1}{C}\sum_{i=1}^{N}\frac{V_{b_i}(\tilde{\mathbf{B}}(\omega)) - V_{c_i}(\tilde{\mathbf{B}}(\omega))}{\sum_{j=1}^{N}\lambda_j(1-\tilde{B}_j(\omega_j))}\cdot\frac{\omega_i}{\sum_{j=1}^{N}(\lambda_j(1-\tilde{B}_j(\omega_j)) - \omega_j)}$$

where $\tilde{\mathbf{B}}(\omega) = (\tilde{B}_1(\omega_1), \ldots, \tilde{B}_N(\omega_N))$.

Please note that the polling system with service disciplines that are not of the branching type defined by Resing [35] cannot be analyzed by exact methods in the general case. Such disciplines are described as follows. If the server arrives at Q_i and finds k_i customers there, then during the server's visit, each of these k_i customers will effectively be replaced in i.i.d. manner by a random population with probability-generating function $h_i(\mathbf{z})$, which can be any N-dimensional probability-generating function. In particular, for the exhaustive service

$$h_i(\mathbf{z}) = \theta_i\left(\sum_{j\neq i}\lambda_j(1-z_j)\right)$$

where $\theta_i(s)$ is the LST of the busy period generated by a customer in queue Q_i which defined as a solution of the functional equation $\theta_i(s) = \tilde{B}_i(s + \lambda_i(1-\theta_i(s)))$. In case of the gated service

$$h_i(\mathbf{z}) = \tilde{B}_i\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right),$$

and for the binomially exhaustive and binomially gated service, respectively,

$$h_i(\mathbf{z}) = (1-p_i)z_i + p_i\theta_i\left(\sum_{j\neq i}\lambda_j(1-z_j)\right), \quad h_i(\mathbf{z}) = (1-p_i)z_i + p_i\tilde{B}_i\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right).$$

The limited-service disciplines are not of the branching type, since for example 1-limited service implies $h_i(\mathbf{z}) = \tilde{B}_i\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right)$ for the first customer at the queue and $h_i(\mathbf{z}) = z_i$ for all other customers. However, some individual cases allow obtaining the generating functions $V_{b_i}(\mathbf{z})$ and $V_{c_i}(\mathbf{z})$ of the joint queue length at the moments of the visit beginning and the server departure Q_i, respectively, e.g., the two-queue system with exhaustive service for Q_1 and k-limited service for Q_2, see Winands et al. [11].

A modified PGF method for the comparative analysis of the individual queue characteristics for an $M/G/1$ polling system was proposed by Guan and Zhao [36]. A queue, say Q_i, is analyzed separately, and the other queues are considered to be the whole queue $Q_{i\sum}$. The further analysis implies derivation of the relations between the PGFs of the number of customers in Q_i and $Q_{i\sum}$ at the polling moments.

Saffer and Telek [37] present the unified analysis for a $BMAP/G/1$ type polling system with exhaustive or gated service. Each queue has $BMAP$ arrival of customers (Batch Markovian Arrival Process, see Dudin et al. [38]) described by the matrix generating function $\hat{D}_i(z) = \sum_{k=0}^{\infty} D_{i,k} z^k$ for queue Q_i, $i = \overline{1,N}$. The methodology of this study is based on the division of the analysis into two parts: dependent and independent of service discipline. Equations are obtained for the system for vector generating functions of the average number of customers in queues, which are valid for a wide class of queuing disciplines and for zero and non-zero switchover times. These equations can be numerically solved as a system of linear algebraic equations.

In the part of the analysis independent of a service discipline, the authors establish the dependence of the vector generating function $\hat{\mathbf{q}}_i(z)$ of the number of customers in queue Q_i on the vector generating functions $\hat{\mathbf{f}}_i(z)$ and $\hat{\mathbf{m}}_i(z)$ of the stationary state distribution of the queue length at the moments the server starts visiting and departs from the queue respectively, within one cycle

$$\hat{\mathbf{q}}_i(z)\hat{D}_i(z)\big(zI - \hat{A}_i(z)\big) = \lambda_i(1-\rho_i^s)(z-1)\frac{\hat{\mathbf{f}}_i(z) - \hat{\mathbf{m}}_i(z)}{f_i^{(1)} - m_i^{(1)}}\hat{A}_i(z)$$

where

$$\hat{A}_i(z) = \int_0^{\infty} e^{\hat{D}_i(z)t} dB_i(t)$$

is the PGF of the number of customers arrived during a customer service time in queue Q_i, $f_i^{(1)}$ ($m_i^{(1)}$) is the mean length of Q_i at the moments the server starts (finishes) service of a customer, $\rho_i^s = \lambda_i^s b_i$ where λ_i^s is the mean intensity of $BMAP$ during the queue service period in Q_i.

The following describes the dynamics of the system in a particular service discipline through multi-dimensional dependent random variables describing the system state at the visit beginnings and completions. A stationary relation

$$\hat{\mathbf{q}}_i^s(z) - \hat{\mathbf{f}}_i^d(z) = \frac{1}{g_i}\hat{\mathbf{f}}_i(z) - \hat{\mathbf{m}}_i(z)$$

is obtained for the vector generating function of the stationary distribution of the number of customers in Q_i where $\hat{\mathbf{q}}_i^s(z)$ ($\hat{\mathbf{q}}_i^d(z)$) is the PGF of the queue length at the service beginning (completion), g_i is the average stationary number of customers served in Q_i during a cycle (it is assumed that the polling order can be non-cyclic).

Van Ommeren et al. [39] consider a polling model with self-ruling service. A server can decide to leave a queue independent of the queue length and the number of served customers. The last service of the server's visit to a queue may differ from the other service times with respect to a service time distribution function. The PGF method is applied to derive the joint probability-generating function of the number of customers and the Laplace transform of the workload in the queues at an arbitrary time.

4.3. Other Methods

This section also includes *the functional calculation approach*. This method was introduced by Hirayama et al. [40]. Its purpose is to find the functional dependencies for the performance characteristics of a cyclic $M/G/1$-type polling system with gated or exhaustive service. Unlike other methods of polling system analysis, this method allows investigation of the system in a transition state. The method considers the expected waiting time of a customer conditioned on the system state at the customer arrival moment. The mean waiting time is considered to be a function of the system states. Furthermore, Hirayama [41] generalizes this model to the case of a Markov feedback and obtain the linear functional relations for the mean waiting times given that the system is in a stationary state. Then Hirayama [42] considers an $M/G/1$-type polling system with a random polling order and a mixed (gated or exhaustive) service discipline.

An analysis of polling systems using *theory of decomposable semiregenerative processes* is described in detail by Rykov [43]. The polling order is supposed to be periodic. The LST of the PGFs of the queue lengths with different types of customers (exhaustive, gated, and limited) are obtained on a separate period of the system performance. Van der Mei [44] proposes the other method to study the polling systems based on the theory of branching processes with migration. The method allows obtaining the approximate expressions for the LSTs of the queue lengths and the waiting times for a wide class of polling systems whose behavior can be described by a branching process.

For the practical application of poling system models, it becomes necessary to numerically implement the methods of polling system analysis to calculate their performance characteristics. Semenova and Bui [45] present a software package to calculate the performance characteristics of polling systems with various types of a polling order (cyclic, adaptive cyclic, random), a wide class of service disciplines (gated, exhaustive, globally gated, limited, threshold and random). The customer input is considered to be Poisson, of the phase type or a correlated MAP. The software package is presented by a simulation module and an analytical calculation module implementing the formulas for calculating the performance characteristics for polling systems that allow the exact analysis. Please note that models of polling systems with correlated input (MAP and $BMAP$) are of high practical importance [38] (for more details on such polling systems, see Section 11). The currently known analytical results of their research for an arbitrary number of queues [37,46] pose an additional problem for the numerical implementation of these results, and this problem is not solved by the authors. Therefore, in the software package, such polling systems can be analyzed by using the simulation.

Vishnevsky et al. [47] propose to use the machine-learning method based on the artificial neural networks to calculate the performance characteristics of polling systems. Machine-learning results are presented for the $M/M/1$ and $MAP/M/1$-type cyclic polling systems and for an $M/M/1$-type system with adaptive cyclic polling. Horng and Lin [48] use machine learning to solve the optimization problem for the limited-service discipline in a $G/G/1/K$-type polling system with k-limited service. Please note that this area of research is new and there are few papers applying machine learning in the field of queuing theory [49,50]. However, as the results of calculations show, the method opens new possibilities to study the polling models which analysis is cumbersome or seems to be impossible by the exact or approximated methods of the theory of random processes.

5. Stability Conditions for Polling Systems

In this section, we note some papers dealing with stability conditions in polling systems. Saffer and Telek [51] establish stability conditions for a system with periodic polling, $BMAP$ input of customers and a mixed service discipline, and generalize the results obtained in their earlier works. They note that there are three possible types of stability of a polling system:

1. *Whole stability*: all queues in the system are stable;
2. *Partial stability* : one or more queues with limited-service discipline are unstable, and the other queues are stable;
3. *Instability*: all queues are instable, and the limiting mean cycle time is infinite.

A queue in a polling system is stable if there exists its stationary state probability distribution (without considering the states of the other queues in the system), and the entire polling system is stable if there exists the stationary state probability distribution at the polling moments and the mean cycle time is finite.

The stability criteria for queue Q_i is $g_i < g_i^{max}$ where $g_i = \lim_{m\to\infty} g_i(m)$ and $g_i(m)$ is the mean number of customers served in queue Q_i at cycle m during all visits (called *stages*) of the server to queue Q_i within this cycle, g_i^{max} is the maximal number of customers the server can serve at the queue during a cycle. For example, $g_i^{max} = \infty$ for the gated service and $g_i^{max} = l$ for l-limited service.

The queue service discipline at the current visit is called unlimited if the average number of customers that can be served during one visit is unlimited given that the queue contains an infinite number of customers at the polling moment (such disciplines include exhaustive, gated, binomial-gated and binomial-exhaustive service). Otherwise, the queue service discipline is called limited (non-exhaustive, semi-exhaustive, k- and T-limited disciplines).

A queue is said to be of unlimited (limited) type if at least at one stage (at all stages) in a cycle, it has unlimited (limited) service discipline. The queue service stage in a cycle means a separate visit period for this queue (starting from the polling moment to the moment the server leaves it). In the case of periodic polling, they can have several visit stages during a cycle.

The limited type queue Q_i is stable if and only if

$$\sum_{k=i}^{N} \rho_i + \frac{\lambda_i}{g_i^{max}} \left(s + \sum_{k=1}^{i-1} g_k^{max} b_k \right) < 1$$

where s is the total mean switchover time of the server during a cycle.

The unlimited type queue Q_i is stable if and only if

$$\rho^U = \sum_{k\in U} \rho_k < 1$$

where U is the set of the numbers of unlimited type queues.

LetQ_1 be the limited type queue then the stability criterion has the form

- whole stability: $\rho + \left(\frac{\lambda_1}{g_1^{max}} r \right) < 1$;
- partial stability: $\rho + \left(\frac{\lambda_1}{g_1^{max}} \right) r \geq 1$ and $\rho^u < 1$;
- instability: $\rho^u \geq 1$;

Also, it is worth noting the paper by Vis et al. [52] analyzing the cycle time of a polling system with gated and globally gated service in non-stationary mode. The first and second moments and the correlation coefficient between two different cycles are obtained (given that the distribution of the first cycle duration is known).

As mentioned in Section 3, Chernova et al.[24] study a three-queue system with limited service by using a fluid model. They show that the stability conditions for such a system cannot be obtained in the closed-form expressions relating to the first moments of the system parameters. They also assume that the stability region may depend on the type of the distributions of service times, switchovers, and interarrival to the queues.

6. Queue Polling Order

This section systematizes the results on the polling system analysis concerning the polling orders. Here we group the papers investigating a random order, a star polling, and cyclic adaptive polling.

6.1. Random Order

Remind that a two-queue system with a random polling order is considered by Dorsman et al. [15]. Then, Dorsman et al. [53] obtain a system of functional equations for the stationary state distribution of an $M/G/1$-type polling system with an arbitrary number of queues.

Hirayama [54] and Fiems and Altman [55] consider the random Markovian polling order (after the server departs from queue Q_i it switches to queue Q_j with probability p_{ij}) and customer feedback. Feedback means that a customer finishing its service can return to the system to repeat its service (see Section 9). In [55], the feedback is described by a semi-linear random process. Hirayama [54] considers the case when a queue can have customers of different priorities. A customer, upon finishing its service, can return to its queue to be served again, change its priority, or transit to another queue. The switchover times are assumed to be zero. In the paper, the various performance characteristics are obtained.

MacPhee et al. [56] consider a polling system with the regeneration of parameters, i.e., every time the server leaves a queue, the queue service, and arrival parameters can change. In this model, only two queues in the system are open at the same time; that is, they are available to accept arriving customers. The other queues do not accept the arriving customers who are lost. Let queue Q_k be open. After the server finishes the queue service, it switches to the next open queue, say Q_i. Then queue Q_k is closed for its arrivals, and the next queue to be opened isQ_j with probability p_{ij}. At the moment the queue service is finished, the arrival and service parameters in open queues are regenerated. The main aim of [56] is to establish the stability conditions for such a system. The case of an arbitrary number of simultaneously opened queues is further investigated by MacPhee at al. in [57]. They consider an $M/M/1$-type polling system with zero switchover times, feedback, and parameter regeneration, as described above. In this regeneration model, not only the customer arrival and service parameters are regenerated, but so do the feedback parameters, i.e., the probabilities that a customer after completing its service stays at the queue. The authors obtain the conditions for any order moments to a system busy period to exist.

Lee [58] considers an $M/PH/1/1$-type polling system with a random polling order and server breakdowns. The server fails only during customer service. It stops service, waits for the repair, and then continues the interrupted customer service.

6.2. Star Polling Order

Remind that for the star polling, the server polls queues in order Q_1, Q_H, Q_2, Q_H, ..., Q_N, Q_H. Guan et al. [59] consider the discrete-time polling system is considered where Q_H is served exhaustively and the other queues get 1-limited service when the server serves no more than one customer from a queue per visit. The analysis provides the mean cycle and the mean queue lengths. Guan and Zhao [36] consider the model of Guan et al. [59] in the case of zero switchover times. The star polling system with gated service for all queues is analyzed by Yang and Ding [60] by the PGF method to obtain the mean number of customers in queues at polling moments.

The star polling system of four $G/G/1$-type queues with exhaustive service of the main queue and 3-phase gated service for other queues is investigated by Bao et al. [61].

6.3. Cyclic Adaptive Polling

Cyclic adaptive polling in polling systems was first considered Vishnevsky et al. [62,63]. With such a polling procedure, the server does not poll a queue (skips it) in the current cycle if the queue was polled in the previous cycle and found empty at the polling moment. We assume that in the case when the server polls N queues sequentially, and all of them are empty (starting from any queue) it stops and takes vacation with a distribution function $H(t)$ with the first and the second moments β and $\beta^{(2)}$ and the LST $\tilde{H}(s)$. As the vacation finishes, the server starts polling the next queue, and the polling procedure is repeated. All queues that the server skips in the current cycle will be polled in the next cycle. Such a model was first analyzed in [62] by decomposition of the polling system into N single-queue systems with server vacations the distribution of which depends on whether the queue was empty at its polling moment or not. The system is studied in a more general form when the input of customers is $BMAP$ (the Batch Markovian Arrival Process), see, e.g., Dudin et al. [38]. Then, the results of [62] are applied to the case of an $M/G/1$-type polling system where the server vacations for a queue are considered to be the time the server spends while visiting the other queues, see Vishnevsky et al. [63]. An approximate algorithm for calculating the main performance characteristics has been developed.

The study of adaptive polling systems is continued by Semenova and Bui [64] and Vishnevsky et al. [65] where the queue length distribution at an arbitrary polling moment is obtained by the PGF method.

The stability condition for an adaptive cyclic polling system is $\rho = \sum_{i=1}^{N} \rho_i < 1$ where $\rho_i = \lambda_i b_i$ is the load of Q_i. The average cycle time for such a system is given by the formula

$$C = \frac{\sum_{i=1}^{N} s_i u_i + \beta \prod_{i=1}^{N} (1 - u_i)}{1 - \rho} \tag{3}$$

where u_i is the probability that Q_i is polled at an arbitrary cycle. This probability is calculated as

$$u_i = \frac{1}{1 + e^{-\lambda_i C}}, \quad i = \overline{1, N}. \tag{4}$$

Equations (3) and (4) provide the system for calculating the mean cycle and probabilities u_i, $i = \overline{1, N}$. Using the PGF method described in detail by Yechiali [33] we get the functional equations for PGFs $F_i(\mathbf{z})$, $\mathbf{z} = (z_1, z_2, ..., z_N)$ of the queue length distributions at the polling moments in case of the gated service:

$$F_i(\mathbf{z}) = u_i \mathbf{M}^{(0)}_{i+1}(\mathbf{z}) + (1 - u_i) u_{i-1} \mathbf{M}^{(1)}_{i+1}(\mathbf{z}) + ... + (1 - u_1) \cdots (1 - u_{N-1}) u_N \mathbf{M}^{(N-1)}_{i+1}(\mathbf{z}) + \\ + (1 - u_1) \cdots (1 - u_N) \mathbf{M}^{(N)}_{i+1}(\mathbf{z})$$

where

$$\mathbf{M}^{(l)}_{i+1}(\mathbf{z}) = F_{i-l}\left(z_1, z_2, ..., z_{i-l}, \widetilde{B}_{i-l}\left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right), z_{i-l+2}, ..., z_N \right) \times \\ \times \widetilde{S}_{i-l+1}\left[\sum_{j=1}^{N} \lambda_j (1 - z_j) \right], \quad l = \overline{0, N-1},$$

$$\mathbf{M}^{(N)}_{i+1}(\mathbf{z}) = F_{i-N}\left(z_1, z_2, ..., z_{i-N}, \widetilde{B}_{i-N}\left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right), z_{i-N+2}, ..., z_N \right) \times \\ \times \widetilde{S}_{i-N+1}\left[\sum_{j=1}^{N} \lambda_j (1 - z_j) \right] \tilde{H}\left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right).$$

And thus, going over the partial derivatives at the point $\mathbf{z}$

$$f_i(j) = \mathbf{M}\left[X_i^j\right] = \left. \frac{\partial F_i(\mathbf{z})}{\partial z_j} \right|_{\mathbf{z}=\mathbf{1}},$$

$$f_i(j,k) = \mathbf{M}\left[X_i^j X_i^k\right] = \left. \frac{\partial^2 F_i(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}}, f_i(i,i) = \mathbf{M}\left[X_i^i (X_i^i - 1)\right] = \left. \frac{\partial^2 F_i(\mathbf{z})}{\partial z_j^2} \right|_{\mathbf{z}=\mathbf{1}}$$

where $\mathbf{1} = (1, ..., 1)$ we get a system of linear equations to calculate the first and second moments of the queue lengths at the polling moments.

Similar results for exhaustive service are obtained by Vishnevsky et al. [65]. The mean waiting time W_i in queue Q_i is calculated as

$$W_i = \frac{f_i(i,i) - f_i}{2 \lambda_i f_i} (1 + \rho_i), i = \overline{1, N},$$

see Yechiali [33] for details.

The similar adaptive polling scheme with skipping the empty queues is considered in a discrete-time polling model by He et al. [4] describing the wireless body area networks, WBANs. In the model, a queue is presented by a wearable wireless sensor or implantable sensor nodes which are used to forward the body signals to a personal server located in intra-WBAN. Each wireless body sensor with no data to transmit and the personal server with no response from all sensors can turn into the inactive state for the energy saving. The PGF method is used to find the stationary state distribution at the moments of sensor polling, while sensors that do not have data to transmit are not polled in a cycle. The average waiting times in the queues, the average cycle time, and other characteristics are obtained.

6.4. Priority Polling

In Section 3, we mentioned that the priority polling order in polling systems is considered by Winands et al. [11], Boon et al. [12], Liu et al. [25] for systems of two or three queues in the system.

The priority in polling systems can be used in the following directions:

1. Using of a polling table (the order to visit the queues) so that the higher-priority queues get more visits by the server in the cycle than the lower-priority queues.
2. Using different queue service disciplines.
3. Varying the customer service order within a queue.

Boon and Adan [66] use a mixed gated and exhaustive service discipline as a combination of the last two prioritization methods. They consider a polling system with $M/G/1$-type queues each with two priority inputs of customers. Priority customers are served exhaustively, and non-priority ones receive the gated service. If the server serves a non-priority customer and a priority customer arrives in the queue, the service is not interrupted, and the server starts serving the priority customer only after the current non-priority service is completed. Thus, this system can be considered to be a system of $2N$ queues $Q_{1H}, Q_{1L}, Q_{2H}, Q_{2L}, ...$ where Q_{iH} are the higher-priority queues and Q_{iL} are the lower-priority queues. It is also assumed that the time to switch between Q_{1H} and Q_{1L} is zero. When the server serves Q_{iL} customers do not arrive to Q_{iH}, i.e., $\lambda^*_{iH} = 0$, and the LST of lower-priority customers in the queue Q^*_{iL} is

$$\beta^*_{iL}(\omega) = \beta_{iL}(\omega + \lambda_{iH}(1 - \pi_{iH}(\omega)))$$

where $\beta_{iL}(\omega)$ the LST of lower-priority customers in Q_i, λ_{iH} is the lower-priority customer arrivals to queue Q_i, $\pi_{iH}(\omega)$ is the busy period distribution for the queuing system corresponding to queue Q_i given that all arrivals are of the higher priority (with parameter λ_{iH}). Thus, the service period of a lower-priority customer in Q_{iL} involves the service times of all higher-priority customers arrived during the service of lower-priority one.

Let the PGFs $V_{b^*_{iH}}(\mathbf{z})$ and $V_{c^*_{iL}}(\mathbf{z})$ be the joint queue length distribution at polling moments of Q_i and at the moments of server's departure from the queue,

$$V_{c^*_{iH}}(\mathbf{z}) = V_{b^*_{iH}}(z_{1H}, z_{1L}, ..., h_{iH}(\mathbf{z}), z_{iL}, ..., z_{NH}, z_{NL})$$

where $\mathbf{z} = (z_{1H}, z_{1L}, ..., z_{NH}, z_{NL})$,

$$h_{iH}(\mathbf{z}) = \pi_{iH}(\alpha_i(\mathbf{z})),\ \alpha(\mathbf{z}) = \lambda_{iL}(1 - z_{iL}) + \sum_{j \neq i}(\lambda_{jH}(1 - z_{jH}) + \lambda_{jL}(1 - z_{jL})),$$

$$V_{c^*_{iL}}(\mathbf{z}) = V_{b^*_{iH}}(z_{1H}, z_{1L}, ..., h_{iH}(\mathbf{z}), h_{iL}(\mathbf{z}), ..., z_{NH}, z_{NL})$$

where $h_{iL}(\mathbf{z}) = \beta^*_{iH}(\alpha_i(\mathbf{z})), \beta^*_{iH}(\omega) = \beta_{iL}(\omega + \lambda_{iH}(1 - \pi_{iH}(\omega)))$.

Please note that $V_{c^*_{iH}}(\cdot) = V_{b^*_{iL}}(\cdot)$ since the switchover times between queues Q^*_{iH} and Q^*_{iL} are zero but switchovers between Q_i and Q_{i+1} are non-zero, thus

$$V_{b^*_{(i+1)H}}(\mathbf{z}) = V_{c^*_{iL}}(\mathbf{z})\sigma_i\left(\sum_{j=1}^{N}(\lambda_{jH}(1-z_{jH}) + \lambda_{jL}(1-z_{jL}))\right) \tag{5}$$

where $\sigma_i(\omega)$ is the LST of the switchover times between queues Q_i and Q_{i+1}.

Boon and Adan [66] note that (5) allows the recursive expression of $V_{b^*_{(i+1)H}}(\mathbf{z})$ through $V_{b^*_{iH}}(\mathbf{z})$ and with further differentiating the derived relations the joint distribution of the queue lengths can be obtained.

The cycle time LST for queue Q_i has the form $\gamma_i(\omega) = \tilde{V}_{b_i}\left(1, 1 - \frac{\omega}{\lambda_{iL}}\right)$ where $\tilde{V}_{b_i}(x, y) = V_{b_i}(1, ..., 1, x, y, 1, ..., 1)$, with x and y are on positions $2i - 1$ and $2i$ corresponding to the higher and lower-priority arrivals to queue Q_i.

The LST of the period when the server does not visit queue Q_i during a cycle (the period from the server's departure from the queue until its next polling moment) has the form $\tilde{I}_i(\omega) = \tilde{V}_{b_i}\left(1 - \frac{\omega}{\lambda_{iH}}, 1\right)$ with the expectation $\mathbf{E}(I_i) = (1 - \rho_i)C$.

The LST of the queue visit time distribution has the form

$$\mathbf{E}\left(e^{-\omega V_i}\right) = \tilde{V}_{b_i}(\pi_{iH}(\omega), \beta^*_{iL}(\omega)).$$

And finally, the LST of the waiting time distribution for Q_i is obtained by

$$\mathbf{E}\left(e^{-\omega W_{iH}}\right) = \frac{(1-\rho_{iH})\omega}{\omega - \lambda_{iH}(1-\beta_{iH}(\omega))}\left[\frac{\rho_{iL}}{1-\rho_{iH}}\frac{1-\beta_{iL}(\omega)}{\omega b_{iL}} + \frac{1-\rho_i}{1-\rho_{iH}}\frac{1-\tilde{I}_i(\omega)}{\omega(1-\rho_i)C}\right]$$

for higher-priority customers, b_{iH} is their mean service time,

$$\mathbf{E}\left(e^{-\omega W_{iL}}\right) = \frac{\tilde{V}_{b_i}(\pi_{iH}(\omega), \beta_{iL}(\omega + \lambda_{iH}(1-\pi_{iH}(\omega)))) - \tilde{V}_{b_i}\left(\pi_{iH}(\omega), 1 - \frac{\omega}{\lambda_{iL}}\right)}{(\omega - \lambda_{iL}(1 - \beta_{iL}(\omega + \lambda_{iH}(1 - \pi_{iH}(\omega)))))C}$$

for lower-priority customers. The paper by Boon et al. [67] generalizes the results of Boon and Adan [66] to the case of general number of the priority levels at each queue.

7. Queue Service Disciplines

Shapira and Levy [68] analyze the fairness of the service disciplines in polling systems (both for discrete polling systems with a finite number of queues and for continuous polling systems). They consider the following disciplines: gated, exhaustive, binomially gated, two-phase gated and globally gated service. For an arbitrary customer, the fairness F is the ratio of the average number of customers served before this customer to the total average number of customers the customer sees in the system upon its arrival. It is shown that the fairest customer service order within a queue is the FIFO order (first come–first-served), in this case $F = 1$. The LIFO (last come–first-served) is the most unfair order, $F = 0$. A random choice of customers for service gives $F = 0.5$. In the case of discrete polling systems, F is calculated by an algorithmic approach described in the paper for the five considered queuing disciplines. It is shown that for polling systems with many queues, the values of F for gated and exhaustive service becomes close, and for the multi-phase gated service $F \to 1$ as the number of phases increases. A globally gated service is fairer than an exhaustive or gated service but less fair than any multi-phase service. The most unfair service order considered is the binomially gated.

In case of a continuous polling system, $F = 2/3$ for both exhaustive and gated service, $F = 8/9$ for two-phase gated service, $F = \frac{6k-4}{6k-3}$ for the k-phase gated service and

$$F = \frac{\sum_{n=0}^{\infty}(1-p)^n\left(\frac{3-p}{12-6p} + \frac{n}{2}\right)}{\sum_{n=0}^{\infty}(1-p)^n(1/2+n)}$$

for the binomially gated service with parameter p, and if $0 < p \leq 1$ we have $1/2 \leq F \leq 2/3$.

Below, we overview the results obtained for the limited service, multi-phase gated service and other service disciplines.

7.1. Limited Service

As noted above, the two-queue $M/M/1$-type polling system with k-limited service is analyzed by Boon and Winands [16]. Van Vuuren and Winands [27] apply the mean value analysis to the $M/G/1$-type polling system with limited-service discipline (see Section 4.1).

Boon et al. [69] detail describe two unsolved problems in the theory of polling systems: analysis of a system with two queues, one with gated service and the other with the 1-limited service, as well as analysis of a system with deterministic, infinitely large switchover times between queues. The paper provides a detailed review in this field and highlights the problems of the limited-service discipline analysis. The latter problem is discussed also by Winands [70] for systems with the branching-type service discipline defined by Resing [35] (see Section 4.2). The paper proposes an asymptotic analysis of the stationary state probability distributions for such systems.

Hanbali et al. [71] consider the model with an autonomous server and T-limited service for an $M_K/G/1$-type polling system with group customer arrivals. Remind that for T-limited service, the server serves the queue until its time has expired (this time is also called a timer) or until the queue is empty whichever occurs first. The autonomous server should stay at a queue until the timer expires regardless of whether the queue is empty or not. These two methods of visiting the queue by the server are not of the branching-type disciplines and this fact makes it difficult to provide an accurate analysis for such systems, see Boxma and Groenendijk [72]. The authors propose an iterative scheme to calculate the joint distributions of the queue length at the polling moments. De Haan et al. [73] analyze two types of preemptive time-limited polling system, the so-called pure and exhaustive time-limited disciplines. They derive a direct relation for the joint queue length behavior during a visit time to a queue. Leonovich and Ferng [74] deal with a T-limited service for an $M/G/1/K$-type polling system.

In paper [46] a $BMAP/G/1$-type polling system with Batch Markovian Arrivals to queues and binomially gated or binomially exhaustive service. The paper generalizes the results of Saffer and Telek [37]. Please note that under the binomially gated service each customer presenting in Q_i at the queue polling moment is marked for service during the current server visit to the queue with probability p_i or is ignored with probability $1 - p_i$ and stays at the queue until the next marking procedure (the next polling moment). Under the binomially exhaustive service, customers are marked in a similar way and the marking procedure also involves customers arriving during the queue service time. Such a system was investigated by the PGF method and functional equations were obtained for the stationary state probabilities at the polling moments and moments when the server leaves a queue. It is noted that these equations can be solved numerically.

De Haan et al. [75] consider an $M/G/1$ -type polling system with T-limited preemptive service. The timer is exponentially distributed. As in Hanbali et al. [71], the server is supposed to be autonomous. It is shown that the stability criteria have the form $\rho_i < \zeta_i$ where

$$\rho_i = \lambda_i \frac{1 - \tilde{B}_i(\xi_i)}{\xi_i \tilde{B}_i(\xi_i)}, \quad \zeta_i = \frac{1/\xi_i}{\sum_{j=1}^{N}(1\xi_j + s_j)},$$

ξ_i is the mean queue Q_i service period, ζ_i is the fraction of time the server spends at queue Q_i during a cycle.

Next the system is investigated by the method of the embedded Markov chains. An approximate analysis of the system is also provided by decomposing it into single-queue systems with the server vacations. Distributions $\omega^i(\mathbf{z})$, $\pi^i_\star(\mathbf{z})$ and $\pi^i(\mathbf{z})$ of the number of customers at the moments of service beginning, service interruption and the successful service completion, respectively, are related as

$$\pi^i(\mathbf{z}) = \frac{1}{\lambda}\tilde{B}_i(\xi_i)\left(\sum_{j=1}^{N}\lambda_j\tilde{B}_j(\xi_j)\right)\check{X}_i((\mathbf{z})\frac{\omega^i(\mathbf{z})}{z_i},$$

$$\pi^i_\star(\mathbf{z}) = (1-\tilde{B}_i(\xi_i))\left(\sum_{j=1}^{N}\frac{\lambda_j}{\tilde{B}_j(\xi_j)}-\lambda_j\right)X^\star_i(\mathbf{z})\omega^i(\mathbf{z})$$

where

$$\check{X}_i((\mathbf{z}) = \frac{\tilde{B}_i(\xi_i+\sum_{j=1}^{N}\lambda_j(1-z_j))}{\tilde{B}_i(\xi_i)},$$

$$X^\star_i(\mathbf{z}) = \frac{\xi_i}{(\xi_i+\sum_{j=1}^{N}\lambda_j(1-z_j)}\cdot\frac{1-\tilde{B}_i(\xi_i+\sum_{j=1}^{N}\lambda_j(1-z_j))}{1-\tilde{B}_i(\xi_i)}.$$

Then, the distributions $a^i(\mathbf{z})$, $b^i_\star(\mathbf{z})$ and $b^i(\mathbf{z})$ of the number of customers at the moments of the server visit beginning, the service preemption and at the end of the server idle time at the queue are related as

$$b^i(\mathbf{z}) = \check{I}_i(\mathbf{z})z_i a^i(\mathbf{z}), \quad b^i_\star(\mathbf{z}) = \check{I}_i(\mathbf{z})a^i(\mathbf{z})$$

where $\check{I}_i(\mathbf{z}) = \frac{\tilde{I}_i(\xi_i+\sum_{j\neq i}\lambda_j(1-z_j))}{\tilde{I}_i(\xi_i)}$, and $\tilde{I}_i(s)$ is the LST of the interarrival time distribution for queue Q_i.

A polling system with $G/G/1/K$-type queues and k-limited service is considered by Horng and Lin [48]. The system parameters are symmetric (independent of the queue number) except the input flow parameters. The authors consider the optimization problem of k_i values (the maximal number of customers that can be served per one visit of the server to the queue) that minimize the cost function (the average cost of waiting for a customer in the system per unit time and the penalty for the loss of customers arriving to the queue if the queue buffer is full). The optimization problem is solved by constructing a neural network, and all performance characteristics are calculated using simulations.

7.2. Multi-Phase Gated Service

The multi-phase gated service as a generalization of the gated discipline was first introduced by van der Mei and Roubos [76]. A customer arriving to queue Q_i must wait k_i cycles before its service, $i = \overline{1,N}$. As the authors note, the goal of such a discipline is to prevent the *monopolization* of the server by more loaded queues by choosing the appropriate levels k_i. The problem is to find the optimal values of k_i for all $i = \overline{1,N}$ minimizing the weighted sum of the mean waiting times in the queues. As noted in the paper, this problem becomes non-trivial in the heavy load conditions and for this case an asymptotic distribution of the waiting times is obtained to calculate the approximate values of the moments and tail of the distribution of waiting times under normal load. Another k_i, $i = \overline{1,N}$ optimization problem is solved by van Wijk [77]. The goal is to maximize the pairwise difference between the average waiting times in queues (the so-called fairness of service) and at the same time maintain efficiency as a weighted sum of average waiting times, i.e., the cost criterion has the form

$$\gamma(\alpha) = (1-\alpha) \max_{i,j=\overline{1,N}} \big(\mathbf{E}[W_i] - \mathbf{E}[W_j]\big) + \alpha \sum_{i=1}^{N} \rho_i \mathbf{E}[W_i]$$

for some parameter $\alpha \in (0,1)$ which allow taking into account the desired relationship between the fairness and the efficiency.

Remerova et al. [78] use the fluid model for a system with multi-phase gated service, an asymptotic analysis of a random process describing the length of an individual. Bao et al. [61] and Ling et al. [79] consider the case of three-phase gated service.

7.3. Other Service Disciplines

Vishnevsky et al. [80] consider a system with the exhaustive threshold polling. A queue can be served if its length exceeds a given threshold. If all queues are not full enough to be served the server stops polling the queues and resumes it as the number of customers at some queue reaches the necessary level. Threshold service disciplines have also been considered for two-queue polling systems by Avrachenkov et al. [20] and Jolles et al. [22] and for tree-queue systems by Liu et al. [25] (see Section 3).

For a multi-dimensional branching process with migration operating in a random environment and producing a final product, Vatutin [81] obtains the tail of the distribution of the product volume produced during the process lifetime. Using this result, the tail distribution of the busy period of polling systems with random polling which are of the branching type, see Resing [35].

8. Queue Scheduling Method

This section lists the papers investigating the various ways to queue (or to group) customers arriving to polling systems. One of methods may include retrial customers. In a retrial system, a customer (initial customer) arriving at the moment when the server is busy goes to the so-called orbit and becomes retrial customer. Then it tries to occupy the server after the random time regardless of other customers in the orbit. If the attempt failed, the customer returns to the orbit again. In the model by Abidini et al. [82] each queue of the polling system is a retrial queue. After the server has switched to the queue it begins a waiting period (or preparation for service) which has a fixed duration. If during this period the primary customer arrives to the queue or a retrial one tries to occupy a place in the queue it is accepted to the queue and at the end of the waiting period the server starts serving customers accumulated in such a way. When all customers accepted to the queue are served, the server leaves the queue. At any other time (out of the server waiting period) the newly arrived customers are not accepted to the queue and become retrials. An addition to Abidini et al. [82] is the paper by Kim and Kim [83] where the PGF method is applied to derive the LST of the waiting time of an arbitrary customer.

Adan et al. [17] consider a symmetric two-queue polling system and arriving customers join the shortest queue.

Next, we note the papers investigating models with group customer service and various methods to form the groups. Dorsman et al. [84] describe a $G/G/1$-type polling system with an internal and external parts where customers (products) can are served by groups. Type i customers arrive to the external system first and are accumulated in a group of type i. As soon as D_i customers are accumulated, the entire group is sent to queue i of the internal system and then it is served as a whole customer. The groups that entered queue Q_i queue in the internal part of the system are then served in the order they were received the next time the server visits the queue. The approximate method by Boon et al. [85] is used to find the weighted sum of the mean waiting times since such a model cannot be analyzed accurately. The problem is to optimize the sizes of groups D_i, $i = \overline{1,N}$ minimizing the weighted sum of the mean waiting times and this problem is solved numerically.

Jiang et al. [86] investigate the polling model where all customers in the queue are served as a group (without limiting the group size). The system operates in a random environment and consists of two areas: service and waiting areas. A customer arriving to the system joins one of the existing groups with some probability or creates a new group (no more than M groups in the system are allowed) and with an additional probability the customer enters the waiting area. Customers in the waiting area form a common queue. As soon as the service of some group ends, the first customer in the waiting area moves to the service area and can form its own group of customers. The system operates in a random multi-phase environment controlled by a Markov chain with a finite state space. If the random environment is in state i then customers arrive from a Poisson input with parameter λ_i and their service times are distributed exponentially with parameter μ_i. The random environment stays in state i during time exponentially distributed with parameter θ_i. Then the environment state can be changed into the neighboring one ($i-1$ or $i+1$). For such a system, the stability conditions are obtained, and the model is analyzed by the matrix-analytical approach to find the stationary state probabilities (see Section 11). The average number of service groups and the LST of the waiting time distribution are obtained.

9. Feedback of Customers

In some system models, it is assumed that at the end of the service, the customers may not leave the system but return to a queue to be served again or go to another queue. As with priority polling systems, a customer can change its priority as, e.g., in Hirayama [54]. This procedure is sometimes called *a feedback of customers*.

A polling system with negative customers is considered by Shomrony and Yechiali [87]. A negative customer is a special type of customers that affects the system performance (remove one or more ordinary customers). For such a system, the LSTs of the queue length distributions and the waiting time distributions are obtained by the PGF method.

Shomrony and Yechiali [88] consider a polling system with two types of breakdowns: a customer service breakdown and a queue service breakdown which can be considered to be negative customers of two types. The first type breakdown happens during a customer service forcing the customer to leave the system partially served. In this case, r = the breakdown does not affect the server which immediately takes the next customer for service. If the server is not connected to the queue the first type breakdown removes the first customer in the queue. The second type breakdown happens during the server's visit to the queue. In this case, the server interrupts its visit, leaves the queue and start switching to the next queue.

Fiems and Altman [55] describe a feedback by a semi-linear random process, and the polling order is random Markovian. They have shown how the behavior of the system can be described using semi-linear stochastic recursive equations in a random Markovian environment and obtain the first and second moments of the number of customers in the system at the polling moments and for the mean number of customers in the system at an arbitrary time. Here a feedback process is described by the amount of work generated by a departing customer rather than the number of customers generated at the departure moment. The service process of a queue is modeled by a semi-linear process.

The stability conditions for an $M/M/1$-type polling system with 1-limited service and a feedback are derived by Zorine [89]. The author assumes that the input flows are controlled by a random environment.

The polling system with impatient customers is considered by Boon [90]. An impatient customer waiting in a queue may leave the system unserved as its waiting time defined by a random variable expires. However, in contrast to most of the queuing systems with impatient customers where customers may leave the system at an arbitrary moment, the model by Boon [90] allows leaving the system only at the moments the server departs from the queue or at the polling moments. A customer leaves the queue with a probability depending on two parameters: the number of the queue and the number of the queue the server is visiting. This way to leave the system is called synchronized, and impatient customers are called *smart* customers. The main difficulty of such a system is that customers leave the system in groups. Using the generalized Little's law in the form of distribution, the distributions of the cycle time, waiting time and the queue lengths are obtained.

A polling system with impatient customers is considered also by Granville and Drekic [91]. The system is presented by two $M/PH/1/b$-type queues with k-limited service. The system states are described by the birth-and-death process, and the stationary state distribution is obtained by using the matrix-analytical approach by Neuts [92] (see Section 11), and the distribution of the waiting time in the system queues is also obtained.

10. Customer Service Order

In this section, we discuss the papers considering how the order of customer service within a queue affect the system performance and the waiting time in particular.

Boxma et al. [93] discuss an $M/G/1$-type polling system with various types of a customer service order within a queue: LCFS (Last Come–First Served, inverse order of service), PS (Processor Sharing), random order, SJF (Shortest Job First, the server chooses the customer with the shortest service time) and others.

The LST of the cycle time distribution for the polling system with a branching-type service discipline [35] is defined as

$$\mathbf{E}\Big(e^{-\omega C}|X_i = m_i, i = \overline{1,N}\Big) = \prod_{i=1}^{N} \sigma(\psi_{i,N}(\omega))\theta_i^{m_i}(\psi_{i,N}(\omega))$$

where X_i is the number of customers in queue Q_i at its arbitrary polling moment, $\theta_i(\omega)$ is the LST of the queue visit time by the server if the queue had a single customer. For the gated service $\theta_i(\omega) = \beta_i(\omega)$, for the exhaustive service $\theta_i(\omega) = \pi_i(\omega)$, i.e., this is the LST of the distribution of the busy period generated by a customer. The functions $\psi_i(\omega)$ and $\psi_{i,j}(\omega)$ are defined as

$$\begin{aligned}
&\psi_i(\omega) = \omega + \lambda_i(1 - \theta_i(\omega)), \quad i = \overline{1,N},\\
&\psi_{i,N}(\omega) = \psi_{i+1}(\psi_{i+2}(...(\psi_N(\omega)))), \quad i = \overline{1,N}, \quad \psi(N,N)(\omega) = \omega.
\end{aligned}$$

The LST of the waiting time distribution in case of gated service has the form

$$\begin{aligned}
\mathbf{E}\Big(e^{-\omega D_{FCFS}}\Big) &= 1 - \mathbf{E}(R_C)(1+\rho_i)\omega + \frac{1}{2}\Big[\lambda_i \mathbf{E}(B_i^2)\mathbf{E}(R_C) + M(R_C^2)(1+\rho_i+\rho_i^2)\Big]\omega^2 + o(\omega^3),\\
\mathbf{E}\Big(e^{-\omega D_{LCFS}}\Big) &= 1 - \mathbf{E}(R_C)(1+\rho_i)\omega + \frac{1}{2}\Big[\lambda_i \mathbf{E}(B_i^2)\mathbf{E}(R_C) + M(R_C^2)(1+\rho_i)^2\Big]\omega^2 + o(\omega^3)
\end{aligned}$$

at $\omega \downarrow 0$, and the first two moments of are calculated by the formulas

$$\begin{aligned}
&\mathbf{E}(D_{FCFS}) = \mathbf{E}(D_{LCFS}) = (1+\rho_i)M(R_C),\\
&\mathbf{E}(D_{FCFS}^2) = \lambda_i \mathbf{E}(B_i)\mathbf{E}(R_C) + \mathbf{E}(R_C^2)(1+\rho_i+\rho_i^2),\\
&\mathbf{E}(D_{LCFS}^2) = \lambda_i \mathbf{E}(B_i)\mathbf{E}(R_C) + \mathbf{E}(R_C^2)(1+\rho_i)^2.
\end{aligned}$$

A random order of service is defined as follows: each customer arriving to a queue is marked by value of a random variable uniformly distributed in [0,1]. Then, at the polling moment all customers waiting for service are arranged in the increasing order of their marks, and the customer with the smallest mark is served first.

The LST of the sojourn time distribution for a customer with mark x in Q_i has the form

$$\mathbf{E}\Big(e^{-\omega T(x)}\Big) = \frac{\beta_1(\omega)}{\omega \mathbf{E}(C)}\Big[\mathbf{E}\Big(e^{-\lambda_i x(1-\beta_1(\omega))C}\Big) - \mathbf{E}\Big(e^{-(\omega+\lambda_i x(1-\beta_1(\omega)))C}\Big)\Big]$$

and the unconditional moments can be computed by integrating out x with respect to a uniform density on [0, 1].

$$\begin{aligned}&\mathbf{E}(T) = \mathbf{E}(B_i) + \mathbf{E}(R_C)(1+\rho_i),\\ &\mathbf{E}(T^2) = \mathbf{E}(T^2_{FCFS}) + \frac{\rho_i}{2} = \mathbf{E}(B_i^2) + \mathbf{E}(R_C)(2(1+\rho_i)\mathbf{E}(B_i) + \lambda_i \mathbf{E}(B_i^2)) + \\ &\quad + \mathbf{E}(R_C^2)\left(1+\rho_i^2+\frac{3}{2}\rho_i\right).\end{aligned}$$

It follows that for the second moments of the mean waiting time for the considered service disciplines it holds

$$\mathbf{E}(T^2_{LCFS}) > \mathbf{E}(T^2_{ROS}) > \mathbf{E}(T^2_{FCFS}).$$

For a general distribution function $B_i(t)$, Processor Sharing and Shortest Job First disciplines make it cumbersome to find the LST of the waiting time distribution in a closed form. However, in case of the exponential distribution, the LST for PS coincides with one for a random service order. Let a customer with length x (its service time) arrives to Q_i then

$$\mathbf{E}(T_{PS}(x)) = x + \mathbf{E}(R_C)[1 + 2\lambda_i \mathbf{E}(min(B_i, x))]$$

and the LST of its waiting time is

$$\begin{aligned}&\mathbf{E}\left(e^{-\omega D_{PS}(x)}\right) = 1 - \mathbf{E}(R_C)(1 + 2\mathbf{E}(min(B_i,x)))\omega + \\ &+ \left[\lambda_i \mathbf{E}(min^2(B_i,x))\mathbf{E}(R_C) + \frac{\mathbf{E}(R_C^2)}{2}\left(1 + 3\mathbf{E}(min(B_i,x)) + 3\mathbf{E}^2(min(B_i,x))\right)\right]\omega^2 + o(\omega^3)\end{aligned}$$

at $\omega \downarrow 0$ then

$$\begin{aligned}&\mathbf{E}(D_{PS}(x)) = \mathbf{E}(R_C)(1 + 2\mathbf{E}(min(B_i,x))),\\ &\mathbf{E}(D_{PS}(x)) = 2\lambda_i \mathbf{E}(min(B_i,x))\mathbf{E}(R_C) + \mathbf{E}(R_C^2)\left(1 + 3\mathbf{E}(min(B_i,x)) + 3\mathbf{E}^2(min(B_i,x))\right).\end{aligned}$$

In case of SJF we have

$$\mathbf{E}\left(e^{-\omega D_{SJF}(x)}\right) = e^{-\omega x}\frac{\mathbf{E}\left(e^{-\lambda_i(1-\phi(\omega,x))C}\right) - \mathbf{E}\left(e^{-(\omega+\lambda_i(1-\phi(\omega,x)))C}\right)}{\omega C}$$

where $\phi(\omega, x) = \mathbf{E}\left(e^{-\omega B_i \mathbf{1}(B_i \le x)}\right)$ and $\mathbf{1}(A)$ is an indicator function of A, $\mathbf{1}(A) = 1$ if A is true and $\mathbf{1}(A) = 0$ otherwise.

Next, the LST of the waiting time distribution is analyzed for a globally gated service. Please note that Boxma et al. [93] consider only gated-type service disciplines which are more convenient for analyzing the system since the sojourn time of customers waiting for service in the current cycle does not depend on customers arriving during the cycle or the server's visit to a queue. In the case of exhaustive service such an analysis seems to be much more complicated.

Bekker et al. [94] study how a customer service order within a queue affects the waiting time distribution in heavy load conditions for the gated and globally gated service. It is known that when customers are served in the order of their arrival, the asymptotic distribution of the waiting time has the form of a product of the uniform and gamma distribution. It is shown that in case of the random service order, the uniform distribution changes to a trapezoidal one, and in case of PS and SJF it changes to a generalized trapezoidal distribution. It is also shown how the choice of the customer service order in each queue affects the behavior of the entire system. Methods to approximate the mean waiting times are developed.

Vis et al. [95] continue the study of Bekker et al. [94] for polling systems with exhaustive service and various types of the customer service order within a queue including LCLS, priority order and others. It is shown that the asymptotic distribution of the cycle time has the form of a gamma distribution with parameters $\alpha = \frac{s\delta}{\sigma^2}$ and $\mu = \frac{\delta}{\sigma^2}$ where $\sigma^2 = \frac{b^{(2)}}{b}$, $\delta = \sum_{i=1}^{N} \hat{\rho}_i(1-\hat{\rho}_i)$, $\hat{\rho}_i$ is the value of load ρ_i at queue i at $\rho = 1$,

$$b^{(k)} = \sum_{i=1}^{N} \frac{\lambda_i b_i^{(k)}}{\lambda}$$

and $b_i^{(k)}$ is the kth moment of the customer service time in queue Q_i.

The LST of the asymptotic distribution of the waiting time in queue Q_i in case of FCFS is

$$W_i^*(x) = \frac{1}{(1-\hat{\rho}_i)sx}\left[1 - \left(\frac{\mu_i}{\mu_i + x}\right)^{\alpha}\right]$$

where $\mu_i = \frac{\delta}{(1-\hat{\rho}_i)\sigma^2}$.

In case of LCLS (both preemptive and non-preemptive service) this distribution has the form

$$W_i^*(x) = \hat{\rho}_i + (1-\hat{\rho}_i)\frac{1}{sx}\left[1 - \left(\frac{\mu}{\mu + x}\right)^{\alpha}\right]$$

where $\mu = \frac{\delta}{\sigma^2}$. The formulas for the other service orders can be found in [95].

Kim and Kim [96] consider an $M/G/1$-type polling system where one of queues has the processor sharing service discipline and the phase-type service time distribution, and customers in other queues are served in FCFS order.

11. Systems with Correlated Input and the Matrix-Analytical Approach

In the present Section, we overview the papers investigating polling systems with correlated arrival of customers. The most commonly used model of a correlated input is a $BMAP$, Batch Markovian Arrival Process, see Dudin et al. [38], allowing description of the properties of the real data streams in modern telecommunication networks. $BMAP$ flow is not stationary, ordinary and the interarrival intervals are correlated. $BMAP$ input is governed by an irreducible non-periodic Markov chain with continuous time and a finite state space. Customers arrive in groups (batches) at the moments the governing process change its states. The transition intensities of the process and a batch size are described by matrices D_k, $k \geq 0$. Matrix D_0 (D_k) defines the transition intensities of the governing process with no customer arrival (with arrival of size k batch of customers, $k \geq 1$). $BMAP$ description has the matrix form and models with $BMAP$ input need to apply a matrix theory and the special matrix theory-based methods to analyze the multi-dimensional stochastic processes necessarily arising while describing the behavior of such systems. The most suitable method in this case seems to be a matrix-analytical approach by Neuts [92]. Below we briefly describe the key points of the method. This approach is used for systems whose states are described by a multi-dimensional Markov random process

$$\xi_t = (i_t, x_{1t}, x_{2t}, ..., x_{Nt}), \quad t \geq 0$$

with $x_{1t}, ', x_{Nt}$ having a finite state space and i_n having finite or denumerable state space, $x_{kt} \in \{1, ..., M_k\}$, $k = \overline{1, N}$, $i_t \geq 0$. The process $\xi_t, t \geq 0$ can be either continuous time or a discrete-time process. Here we consider the case of a continuous time process. The states of process $x_{1n}, x_{2n}, ..., x_{Nn}$ are numbered in lexicographical order and, thus, we get a two-dimensional process $\tilde{\xi}_n = (i_n, \tilde{x}_n)$, $n \geq 0$ where $\tilde{x}_n \in \{1, ..., \prod_{i=1}^{N} M_i\}$ and its infinitesimal generator $\tilde{Q}$ has a block structure

$$\tilde{Q} = \begin{pmatrix} B_0 & A_0 & & & & \\ B_1 & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \dots \\ & & & A_2 & A_1 & \dots \\ & & & \vdots & \vdots & \end{pmatrix}.$$

It is assumed that the structure of $\tilde{Q}$ is irreducible and $(B_0 + A_0)\mathbf{1} = (B_1 + A_1 + A_0)\mathbf{1} = (A_0 + A_1 + A_2)\mathbf{1} = \mathbf{0}$ where $\mathbf{1}$ is a column vector consisting of 1's, $\mathbf{0}$ is a zero row vector.

Then let **ß** be the vector of the stationary state probabilities for generator $A = A_0 + A_1 + A_2$, i.e., $\text{ß}A = \mathbf{0}$. Then the process $\tilde{\xi}_n$, $n \geq 1$ has a stationary state distribution $\mathbf{x} = \{\mathbf{x}_0, \mathbf{x}_1, ...\}$ if and only if

$$\text{ß}A_2\mathbf{1} > \text{ß}A_0\mathbf{1}.$$

The ith entry of vector $\mathbf{x}$ is vector $\mathbf{x}_i$ of the stationary state distribution of the process $\tilde{\xi}_n = (i_n, \tilde{x}_n)$, $n \geq 0$ given that $i_n = i$. Vectors $\mathbf{x}_i$, $i \geq 1$ are obtained in the form

$$\mathbf{x}_i = \mathbf{x}_0 R^i, \quad i \geq 1$$

where matrix R is the minimal nonnegative solution of the matrix equation $R^2 A_2 + R A_1 + A_0 = O$ and $\mathbf{x}_0$ is the unique solution of the system

$$\mathbf{x}_0(B_0 + RB_1) = \mathbf{0},$$
$$\mathbf{x}_0(I - R)^{-1}\mathbf{1} = 1.$$

For systems whose behavior is described by the random processes with infinitesimal generators of a more complex form (for example, when the block structure depends on the level i) there are modifications of the matrix-analytical approach which can be found in Dudin et al. [38] in more details.

Remind that a generalized analysis of a $BMAP/G/1$-type polling system with a gated or exhaustive service is presented by Saffer and Telek [37] (see Section 4.2) where the relations for the vector generating functions of the mean queue lengths are obtained which are valid for a wide class of queuing disciplines and for both zero and non-zero switchover times. These equations can be numerically solved as a system of linear algebraic equations.

Cao et al. [3] consider a system of two $BMAP/PH/1$-type queues with a limited number of waiting space. The input flows are $BMAP$s and the customer service time distribution is of the phase type. This model describes the features of modern video compression standards. The first queue gets the gated service and the second one has T-limited service depending on the state of Q_1 at the moment the server departs from Q_1. The joint queue length probability distribution is obtained by using the embedded Markov chain method and the matrix-analytical approach described above. The stability conditions for such system in the case of an unlimited number of waiting places in queues are obtained by Cao and Xie [97].

Chen [98] considers polling systems where customers arrive in groups and simultaneously to all queues. The interarrival time is exponentially distributed and the size of customer groups is defined by the random vector $K = (k_1, k_2, .., k_N)$ with a probability-generating function $\mathbf{k}(z) = k(z_1, z_2, ...z_N)$. Since the average waiting time in queues depends on the type of distribution functions of service time, interarrival time and server switchover times only through their first two moments, the author suggests applying the pseudo-transformation of the initial distributions by the generating function of moments which allows reducing the number of required calculations in comparison with the classical PGF method for polling systems.

We also note here the papers investigating polling models with uncorrelated input but using a matrix-analytical approach to find the stationary state probability distributions. More details on these models a reader can find in other sections of the review. Jolles et al. [22] apply a matrix-analytical approach for the two-queue system with a threshold server switching strategy. Perel et al. [23] consider a system where a server chooses the longest queue to serve.

Jiang et al. [86] consider the polling system where all customers in a queue are served as a whole group (with no limits on the group size). The system operates in a random multi-phase environment controlled by a Markov chain with a finite state space that determines the parameters of input flows and service. Granville and Drekic [91] apply the matrix-analytical approach to the analysis of an $M/PH/1/b$-type polling system with two queues, k-limited service and impatient customers. Suman and Krishnamurthy [99] analyze a tandem of two $M/M/1$-type polling systems with two queues.

12. Multi-Server Polling Systems

Antunes et al. [100] consider a multi-server polling system assuming that the number of servers that can simultaneously visit the same queue is limited. Servers visit the queues independently of each other in a random order. For such a system, the stability conditions for a system with unlimited type queue service disciplines are obtained. Recall that the service discipline is of unlimited type if the number of customers the server can serve during one visit to a queue is unlimited. We should also note the paper by Vishnevsky et al. [32] studying the polling model with a duplex queue polling where the queues are polled by two independent servers. Some of the queues are common for both servers and the others are assigned to one of servers. The system analysis is based on the mean value analysis.

Boxma et al. [101] investigate a polling system with an unlimited server resource (all customers waiting in the queue at the polling moment are simultaneously served). Such a system can also be represented as a system with an unlimited number of servers visiting the queues simultaneously. The system is analyzed by the PGF method to obtain the waiting time distribution. A similar system with T-limited service was investigated in [102]. As the timer of the server to visit a queue expires the service is interrupted and customers on servers will be served again at the next visit to the queue. For such a system, the LST of the sojourn time distribution is obtained.

13. Polling Systems with Heavy Traffic

This section presents the main results on the polling system analysis in heavy load conditions when $\rho \to 1$. For some polling systems, it is possible to obtain the approximate formulas to calculate the performance characteristics by considering ρ as a variable, and the parameters describing the system are presented as the functions of ρ.

Van der Mei and Winands [103] analyze a polling system with gated service where the input of customers to a queue is represented by a recovery process with parameters λ_i, $i = \overline{1,N}$. The load ρ is increased in such a way that only the input flow parameters increase, and the service time distribution and the fraction of the queue load share $\frac{\rho_i}{\rho}$ keep the same. As the system load ρ increases, all queues in the system become unstable, and the average waiting times $\mathbf{E}[W_i]$ in queues tend to infinity for all $i = \overline{1,N}$. Thus, $\mathbf{E}[W_i]$ as a function of ρ has a first-order pole at the point $\rho = 1$ (see van der Mei and Levi [104]), i.e.,

$$\mathbf{E}[W_i] = \frac{\omega_i}{1-\rho} + o((1-\rho)^{-1}), \qquad \rho \uparrow 1, i = \overline{1,N}$$

where ω_i is the mean asymptotically normalized waiting time in queue Q_i, the value ω_i is the rate for $\mathbf{E}[W_i]$ to grow infinitely as $\rho \uparrow 1$.

Parameters ω_i are defined as

$$\omega_i = \frac{(1+\hat{\rho}_i)}{2}\left(\frac{\sigma^2}{\sum_{i=1}^N \hat{\rho}_j(1+\hat{\rho}_j)} + r\right)$$

where $\sigma^2 = \sum_{i=1}^N \hat{\lambda}_i(\mathbf{D}[B_i] + \hat{\rho}_i^2\mathbf{D}[\hat{A}_i])$, $\mathbf{D}[B_i]$ is the service time variance, $\hat{\lambda}_i = \frac{1}{\mathbf{E}[\hat{A}_i]}$ and $\mathbf{D}[\hat{A}_i]$ are the parameter and variance of the queue Q_i interarrival times at $\rho = 1$, respectively, $\hat{\rho}_i = \hat{\lambda}_i b_i$.

Dorsman et al. [105] consider a polling system with the arbitrary interarrival time distribution functions under heavy load conditions and analyze by using the results of Boon et al. [85]. A $G/G/1$-type polling system in heavy load conditions was considered by Boon et al. [106] and analyzed by using the corresponding fluid model. Van der Mei and Winands [28] propose a new approach to the approximate analysis of polling systems with gated or exhaustive service in heavy load based on the mean value analysis by Winands et al. [26] which allow obtaining the linear system of equations for the mean values $L_{i,j}$ of the number of customers in Q_i at an arbitrary time when the server visits queue Q_j. Considering the system in heavy load as $\rho \to 1$, all parameters describing the system (input flows, service and switchover times) are considered to be functions of ρ. The value of parameter x at $\rho = 1$ is denoted as $\hat{x}$. Let also $\mathbf{M}[L_{i,j}^*] = \lim_{\rho\uparrow 1}(1-\rho)\mathbf{M}[L_{i,j}]$ then for the exhaustive service it holds

$$\begin{aligned}
&\mathbf{M}[L_{i,i}^*] = c\hat{\lambda}_i(1-\hat{\rho}_i),\\
&\mathbf{M}[L_{i,i+n}^*] = c\hat{\lambda}_i\left(2\sum_{m=1}^{n-1}\hat{\rho}_{i+m} + \hat{\rho}_{i+n}\right), \quad i = \overline{1,N}, n = \overline{1,N-1}
\end{aligned}$$

where $c = \frac{1+\beta\delta s}{2\beta\delta}$, $\beta = \frac{b}{b^{(2)}}$, $\delta = 1 - \sum_{i=1}^N \hat{\rho}_i^2$ and for the exhaustive service it holds

$$\begin{aligned}
&\mathbf{M}[\hat{L}_{i,i}^*] = c\hat{\lambda}_i, \quad \mathbf{M}[\tilde{L}_{i,i}^*] = c\hat{\lambda}_i\hat{\rho}_i,\\
&\mathbf{M}[\tilde{L}_{i,i+n}^*] = c\hat{\lambda}_i\left(2\sum_{m=1}^{n-1}\hat{\rho}_{i+m} + \hat{\rho}_{i+n}\right), \quad i = \overline{1,N}, n = \overline{1,N-1}
\end{aligned}$$

where $\hat{L}_{i,i}$ and $\tilde{L}_{i,i}$ mean the queue length before the gate (customers to be served during the server's visit to Q_i) and behind the gate (customers to be served at the next visit), respectively.

The asymptotical mean values (scaled for $1-\rho$) of the waiting times are given by $\mathbf{M}[W_i^*] = c(1-\hat{\rho}_i)$ for exhaustive service and by $\mathbf{M}[W_i^*] = c(1+\hat{\rho}_i)$ for gated service, respectively.

Paper [44] shows that for $X_{i,i+n}$ ($Y_{i,i+n}$) denoting the number of customers in queue Q_i at the beginning (at the end) of an arbitrary visit period of the server to queue Q_{i+n} it holds

$$\begin{aligned}
&(1-\rho X_{i,i+n}) \to_d d\hat{\lambda}_i\left(\sum_{m=1}^{n-1}\hat{\rho}_{i+m}\right)\Gamma(\alpha,1),\\
&(1-\rho Y_{i,i+n}) \to_d d\hat{\lambda}_i\left(\sum_{m=1}^{n-1}\hat{\rho}_{i+m} + \hat{\rho}_{i+n}\right)\Gamma(\alpha,1), \quad i = \overline{1,N}, n = \overline{1,N-1}
\end{aligned}$$

where $\to_d$ is the convergence in distribution, d is the known constant, $\Gamma(\alpha,1)$ is Gamma distribution with parameter α independent of i and $i+n$.

Remind that a system with two $M/M/1$-type queues and a limited-service discipline in heavy load conditions was investigated by Boon and Winands [16] (see Section 3). This section also includes the paper by Meyfroyt et al. [107] considering a symmetric polling system with many queues and branching-type service disciplines [35]. It is shown that as the number of queues grows infinitely large and the total system load keeps the same value, the random variables describing the queue lengths become independent in the limit, and this shows how the behavior of a single queue can be described in terms of the single-queue system with the server vacations which simplifies the analysis of the queue length and waiting time distributions. A flexible k-limited-service discipline has also been introduced to reduce the mean queue lengths and the mean cycle time for delay-sensitive applications.

14. Non-Discrete Polling Systems and Polling Networks

Non-discrete polling models (networks) means systems where arriving customers are placed on a circle, systems with denumerable number of queuing places and the fluid.

Kavitha and Combes [108] provides a generalized analysis of discrete polling systems (with a finite number of queues) and continuous systems. Such systems in the paper are called mixed polling systems. It is also assumed that customers, having finished their service, can move to other queues (or change their location in the case of a continuous polling system). For a continuous system, the authors present the method of the system discretization and obtain the average amount of virtual work, i.e., the average time that the server will work while serving all customers in the system at a given time, as well as the residual service time if the server is currently busy.

Boxma et al. [109] consider a cyclic polling system with a mixed service discipline in which queue load (the amount of work that needs to be processed by the server) is controlled by a positively incremented N-dimensional Levy process. Such a process means the dependence of incoming flows in the queue. A stationary distribution of the amount of work in the system at embedded and at arbitrary moments of time is obtained.

Leskela and Unger [110] and Kavitha and Altman [111] consider models where arriving customers are placed on a circle. Leskela and Unger [110] assume that the server serves customers located in the vicinity of the server's current position, and the server selects the closest customer to serve. The classical stability condition where the service rate must exceed the arrival rate is proved by Kavitha and Altman [111]. For the system considered, the server moves along a circle at a constant speed and some customers are served according to the globally gated discipline, and the others are served according to the exhaustive rule. The main idea of the paper is to use a discrete polling model with a finite number of queues in which the average virtual load of the system is obtained as the limit of the average virtual load of a discrete polling system with a mixed service discipline for which the pseudo-conservation law can be applied.

We also mention here paper by Beekhuizen et al. [112] analyzing the polling network. Polling networks consist of several polling systems and customers transit between nodes. Having received service in a queue of one node, a customer moves to another node (another polling system), joins a queue and waits for its service. The polling network has a tree structure where customers served at a network node moves to the neighboring node and so on until it reaches the root node and then leaves the network. It is shown how the behavior of a polling network can be described in terms of a single (root) node.

The fluid model of the polling system is presented by Matveev et al. [113]. A queue is interpreted as the fluid level which decreases when the server serves the queue. The service speed can vary depending on the liquid level according to the selected speed control strategy.

Saffer et al. [114] discuss a fluid model of a gated service polling system. The increase in the amount of work in the queue is controlled by a continuous time Markovian process and it decreases at a constant speed when the queue is served. The necessary conditions for the existence of a stationary regime are obtained, as well as relations for the Laplace vector transformations of stationary fluid levels in the system queues at the polling moments, the moments the server leaves the queue, and at an arbitrary time, as well as other stationary distributions that characterize the behavior of the system. It is shown that the average cycle time for such a model has the classical form $C = s/(1-\rho)$.

In Yechiali and Czerniak [115], the polling system consists of N fluid-queuing systems and a single server. Service disciplines are exhaustive, gated or globally gated. The order of polling queues is cyclic or random. The LST of the fluid level distribution in the queues at the time of the server polls the queues and at an arbitrary moment is obtained. In addition, the procedure for finding the optimal probabilistic order of polling queues is described.

For a system with multi-phase gated service Remerova et al. [78] provide an asymptotic analysis of a random process that describes the length of a single queue using a fluid model. A feedback $G/G/1$-type polling system is considered by Boon et al. [106]. Chernova et al. [24] obtain the stability condition for a system with three queues and adaptive polling. Czerniak et al. [116] propose an analysis of the polling system with queues represented by fluid models with cyclic and random polling order to describe the TCP protocol.

A tandem of two polling systems each consisting of two $M/M/1$-type queues with cyclic polling and exhaustive service is studied by Suman and Krishnamurthy [99]. After completing the service in the first system, a customer moves to the corresponding queue of the second tandem system. The following policies for servicing queues are considered: servers cyclically serve queues independently of each other, synchronously switch to the queues with the same numbers (one of the servers may be idle) or synchronously switch to queues with different numbers. The paper investigates the impact of these service policies on waiting time using a matrix-analytical approach.

15. Conclusions

In the paper, we proposed a review of papers published from 2007–2020 in the field of polling systems. In contrast to the recent review by Borst and Boxma [10] where the main focus is placed on cyclic polling systems with a single server and new approximations for systems in heavy load and systems with a large number of queues, we tried to provide the possibly full range of models and methods presented in the recent literature in the field of polling systems. We described the directions of theoretical research development in the area and pointed out new practical applications of polling systems. We also noted some unsolved theoretical problems and proposed to apply the machine-learning method to solve new polling models which analysis is cumbersome or imply no closed-form solution.

Author Contributions: Conceptualization, V.V. and O.S.; validation, V.V. and O.S.; investigation, V.V. and O.S.; writing–original draft preparation, O.S.; writing–review and editing, V.V.; supervision, V.V.; funding acquisition, V.V. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by the Russian Foundation for Basic Research, project no. 19-29-06043.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vishnevsky, V.; Semenova, O. *Polling Systems: Theory and Applications for Broadband Wireless Networks*; LAMBERT Academic Publishing, Saarbrücken, Germany: 2012; 317p.
2. Boon, M.A.A.; van der Mei, R.D.; Winands, E.M.M. Applications of polling systems. *Surv. Oper. Res. Manag.* **2011**, *16*, 67–82.
3. Cao, J.; Feng, W.; Chen, Y.; Ge, N.; Wang, S. Performance analysis of a polling model with BMAP and across-queue state-dependent service discipline. *IEEE Access* **2019**, *7*, 127230–127253.
4. He, M.; Guan, Z.; Wu, Z.; Lu, L.; Zhou, Z.; Anisetti, M.; Damiani, E. A polling access control with exhaustive service in wireless body area networks for mobile healthcare using the sleeping schema. *J. Ambient. Intell. Humaniz. Comput.* **2019**,*10*,3761–3774.
5. Granville, K.; Drekic, S. A 2-Class Maintenance Model with Dynamic Server Behavior. *TOP* **2019**, doi:10.1007/s11750-019-00509-1
6. Takagi, H. *Analysis of Polling Systems*; MIT Press: Cambridge, MA, USA, 1986.
7. Borst, S.C. *Polling Systems*; Stichting Mathematisch Centrum: Amsterdam, The Netherlands, 1996.
8. Vishnevskii, V.M.; Semenova, O.V. Mathematical methods to study the polling systems. *Autom. Remote Control* **2006**, *67*, 173–220.

9. Vishnevsky, V.M.; Mishkoy, G.K.; Semenova, O.V. New models and methods to study polling systems. In Proceedings of the International Conference proceedings Distributed Computer and Communication Networks. Theory and Applications (DCCN-2009), Sofia, Bulgaria, 5–9 October 2009; pp. 79–85. (In Russian)
10. Borst, S.C.; Boxma, O.J. Polling: past, present, and perspective. *TOP* **2018**, *26*, 335–369.
11. Winands, E.M.M.; Adan, I.J.B.F.; van Houtum, G.J.; Down, D.G. A state-dependent polling model with *k*-limited service. *Probab. Eng. Infor. Sci.* **2009**, *23*, 385–408.
12. Boon, M.A.A.; Adan, I.J.B.F.; Boxma, O.J. A two-queue polling model with two priority levels in the first queue. *Discret. Event Dyn. Syst.* **2010**, **20**, 511–536.
13. Vlasiou, M.; Adan, I.J.B.F.; Boxma, O.J. A two-station queue with dependent preparation and service times. *Eur. J. Oper. Res.* **2009**, *195*, 104–116.
14. Chernova, N.; Foss, S.; Kim, B. A polling system whose stability region depends on the whole distribution of service times. *Oper. Res. Lett.* **2013**, *41*, 188–190.
15. Dorsman, J.-P.L.; Boxma, O.J.; van der Mei, R.D. On two-queue Markovian polling systems with exhaustive service. *Queueing Syst.* **2014**, *78*, 287–311.
16. Boon, M.A.A.; Winands, E.M.M. Heavy-traffic analysis of *k*-limited polling systems. *Probab. Eng. Informational Sci.* **2014**, *28*, 451–471.
17. Adan, I.J.B.F.; Boxma, O.J.; Kapodistria, S.; Kulkarni, V.G. The shorter queue polling model. *Ann. Oper. Res.* **2016**, *241*, 167–200.
18. Gaidamaka, Yu.V. Model with threshold control for analysing a server with SIP protocol in the overload mode. *Autom. Control. Comput. Sci.* **2013**, *47*, 211–218.
19. Shorgin, S.; Samouylov, K.; Gaidamaka, Y.; Etezov, S. Polling system with threshold control for modeling of SIP server under overload. *Adv. Intell. Syst. Comput.* **2014**, *240*, 97–107.
20. Avrachenkov, K.; Perel, E.; Yechiali, U. Finite-buffer polling system with threshold-based switching policy. *TOP* **2016**, *24*, 541–571.
21. Perel, E.; Yechiali, U. Two-queue polling systems with switching policy based on the queue that is not being served. *Stoch. Model.* **2017**, *33*, 430–450.
22. Jolles, A.; Perel, E.; Yechiali, U. Alternating server with non-zero switch-over times and opposite-queue threshold-based switching policy. *Perform. Eval.* **2018**, *126*, 22–38.
23. Perel, E.; Perel, N.; Yechiali, U. A polling system with «Join the shortest – serve the longest» policy. *Comput. Oper. Res.* **2020**, *114*, 104809.
24. Chernova, N.; Foss, S.; Kim, B. On the stability of a polling system with an adaptive service mechanism. *Ann. Oper.* **2012**, *198*, 125–144.
25. Liu, Z.; Chu, Y.; Wu, J. On the three-queue priority polling system with threshold service policy. *J. Appl. Math. Comput.* **2017**, *53*, 445–470.
26. Winands, E.M.M.; Adan, I.J.B.F.; van Houtum, G.J. Mean value analysis for polling systems. *Queueing Syst.* **2006**, *54*, 35–44.
27. van Vuuren, M.; Winands, E.M.M. Iterative approximation of *k*-limited polling systems. *Queueing Syst.* **2007**, *55*, 161–178.
28. van der Mei, R.D.; Winands, E. Heavy traffic analysis of polling models by mean value analysis. *Perform. Eval.* **2008**, *65*, 400–416.
29. Vishnevsky, V.M.; Semenova, O.V. Adaptive dynamical polling in wireless networks. *Cybern. Inf. Technol.* **2008**, *8*, 3–11.
30. Wierman, A.; Winands, E.; and Boxma, O.J. Scheduling in polling systems. *Perform. Eval.* **2007**, *64*, 1009–1028.
31. Boon, M.A.A.; van Wijk, A.C.C.; Adan, I.J.B.F.; Boxma, O.J. A polling model with smart customers. *Queueing Syst.* **2010**, *66*, 239-274.
32. Vishnevskii, V.M.; Semenova, O.V.; Shpilev, S.A. A duplex cyclic polling system for mixed queues. *Autom. Remote Control* **2009**, *70*, 2050–2060.
33. Yechiali, U. Analysis and control of polling systems. In *Performance Evaluats of Computer and Communication Systems*; Donatielo, L., Nelson, R.; Eds.; Springer: Berlin/Heidelberg, Germany, 1993; pp. 630–650.
34. Boxma, O.J.; Kella, O.; Kosinski, K.M. Queue lengths and workloads in polling systems. *Oper. Res. Lett.* **2011**, *39*, 401–405.
35. Resing, J.A.C. Polling systems and multitype branching processes. *Queueing Syst.* **1993**, *13*, 413–426.
36. Guan, Z.; Zhao, D. A delay-guaranteed two-level polling model. *Adv. Comput. Sci. Inf. Eng. Adv. Intell. Soft Comput.* **2012**, *168*, 153–158.
37. Saffer, Z.; Telek, M. Unified analysis of $BMAP/G/1$ cyclic polling models. *Queueing Syst.* **2010**, *64*, 69–102.
38. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *Methods to Study Queuing Systems with Correlated Arrivals*; Springer: Berlin/Heidelberg, Germany, 2020; 410p.
39. van Ommeren, J.-K.; Hanbali, A.A, Boucherie, R.J. Analysis of polling models with a self-ruling server. *Queueing Syst.* **2020**, *94*, 77–107.
40. Hirayama, T.; Hong, S.J.; Krunz, M.M. A new approach to analysis of polling systems. *Queueing Syst.* **2004**, *48*, 135–158.
41. Hirayama, T. Multiclass polling systems with Markovian feedback: mean sojourn times in gated and exhaustive systems with local priority and FCFS service orders. *J. Oper. Res. Soc. Jpn.* **2005**, *48*, 226–255.
42. Hirayama, T. Markovian polling systems: functional computation for mean waiting times and its computational complexity. In *Advances in Queueing Theory and Network Applications*; Yue, W., Ed.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 119–146.
43. Rykov, V.V. On analysis of periodic polling systems. *Autom. Remote Control* **2009**, *70*, 997–1018.
44. van der Mei, R.D. Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Syst.* **2007**, *57*, 29–46.

45. Semenova, O.V.; Bui, D.T. The software package and its application to study the polling systems. *Vestn. Tomsk. Gos. Univ. Upr. Vychislitelnaja Teh. Inform. [Tomsk. State Univ. J. Control. Comput. Sci.]* **2020**, *50*, 106–113. (In Russian)
46. Saffer, Z. $BMAP/G/1$ cyclic polling model with binomial disciplines. *Mod. Probabilistic Methods Anal. Telecommun. Commun. Comput. Inf. Sci.* **2013**, *356*, 157–166.
47. Vishnevsky, V.M.; Semenova, O.V.; Bui, D.T. Using machine learning to study polling systems with correlated flow input. In Proceedings of the Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems (ITTMM 2020); RUDN, Moscow, Russia, 13–17 April 2020; pp. 248–253. (In Russian)
48. Horng, S.-C.; Lin, S.-Y. Ordinal optimization of $G/G/1/K$ polling systems with k-limited service discipline. *J. Optim. Theory Appl.* **2009**, *140*, 213–231.
49. Gorbunova, A.V.; Vishnevsky, V.M. Estimating the response time of a cloud computing system with the help of neural networks. *Adv. Syst. Sci. Appl.* **2020**, *20*, 105–112.
50. Larionov, A.; Vishnevsky, V.; Semenova, O.; Dudin, A. A multiphase queueing model for performance analysis of a multi-hop IEEE 802.11 wireless network with DCF channel access. *Commun. Comput. Inf. Sci.* **2019**, *1109*, 162–176.
51. Saffer, Z.; Telek, M. Stability of periodic polling system with BMAP arrivals. *Eur. J. Oper. Res.* **2009**, *197*, 188–195.
52. Vis, P.; Bekker, R.; van der Mei, R.D. Transient analysis of cycle lengths in cyclic polling systems. *Perform. Eval.* **2015**, *91*, 303–317.
53. Dorsman, J.-P.L.; Borst, S.C.; Boxma, O.J.; Vlasiou, M. Markovian polling systems with an application to wireless random-access networks. *Perform. Eval.* **2015**, *85–86*, 33–51.
54. Hirayama, T. Analysis of multiclass Markovian polling systems with feedback and composite scheduling algorithms. *Ann. Oper. Res.* **2012**, *198*, 83–123.
55. Fiems, D.; Altman, E. Gated polling with stationary ergodic walking times, Markovian routing and random feedback. *Ann. Oper. Res.* **2012**, *198*, 145–164.
56. MacPhee, I.; Menshikov, M.; Petritis, D.; Popov, S. A Markov chain model of a polling system with parameter regeneration. *Ann. Appl. Probab.* **2007**, *17*, 1447–1473.
57. MacPhee, I.; Menshikov, M.; Petritis, D.; Popov, S. Polling systems with parameter regeneration, the general case. *Ann. Appl. Probab.* **2008**, *18*, 2131–2155.
58. Lee, T. Analysis of single buffer random polling system with state-dependent input process and server/station breakdowns. *Int. J. Oper. Res. Inf. Syst. IJORIS* **2018**, *9*, 22–50.
59. Guan, Z.; Zhao, D.; Zhao, Y. A discrete time two-level mixed service parallel polling model. *J. Electron. China* **2012**, *29*, 103–110.
60. Yang, Z.; Ding, H. Characteristics of a two-class polling system model. *Tsinghua Sci. Technol.* **2014**, *19*, 516–520.
61. Bao, L.; Zhao, D.; Zhao, Y. A priority-based polling scheduling algorithm for arbitration policy in Network on Chip. *J. Electron. China* **2012**, *29*, 120–127.
62. Vishnevsky, V.M.; Dudin, A.N.; Semenova, O.V.; Klimenok, V.I. Performance analysis of the $BMAP/G/1$ queue with gated servicing and adaptive vacations. *Perform. Eval.* **2011**, *68*, 446–462.
63. Vishnevsky, V.M.; Dudin, A.N.; Klimenok, V.I.; Semenova, O.V.; Shpilev, S. Approximate method to study $M/G/1$-type polling system with adaptive polling mechanism. *Qual. Technol. Quant. Manag.* **2012**, *2*, 211–228.
64. Semenova, O.V.; Bui, D.T. Method of generating functions for performance characteristic analysis of the polling systems with adaptive polling and gated service. *Commun. Comput. Inf. Sci.* **2018**, *912*, 348–359.
65. Vishnevsky, V.M.; Semenova, O.V.; Bui, D.T.; Sokolov, A.M. Adaptive cyclic polling systems: analysis and application to the broadband wireless networks. *Lect. Notes Comput. Sci.* **2019**, *11965*, 30–42.
66. Boon, M.A.A.; Adan, I.J.B.F. Mixed gated/exhaustive service in a polling model with priorities. *Queueing Syst.* **2009**, *63*, 383–399.
67. Boon, M.A.A.; Adan, I.J.B.F.; Boxma, O.J. A polling model with multiple priority levels. *Perform. Eval.* **2010**, *67*, 468–484.
68. Shapira, G.; Levy, H. On fairness in polling systems. *Ann. Oper. Res.* **2016**, doi:10.1007/s10479-016-2247-8
69. Boon, M.; Boxma, O.J.; Winands, E.J.J. On open problem in polling systems. *Queueing Syst.* **2011**, *68*, 365–374.
70. Winands, E.M.M. Branching-type polling systems with large setups. *OR Spectr.* **2011**, *33*, 77–97.
71. Hanbali, A.A.; de Haan, R.; Boucherie, R.J.; van Ommeren, J.-K. Time-limited polling systems with batch arrivals and phase-type service times. *Ann. Oper. Res.* **2012**, *198*, 57–82.
72. Boxma, O.J.; Groenendijk, W.P. Pseudo conservation laws in cyclic-service systems. *J. Appl. Probab.* **1987**, *24*, 949–964.
73. de Haan, R.; Hanbali, A.A.; Boucherie, R.J.; van Ommeren J.-K. Transient analysis for exponential time-limited polling models under the preemptive repeat random policy. *Adv. Appl. Probab.* **2020**, *52*, 32–60.
74. Leonovich, A.; Ferng, H.-W. Modeling the IEEE 802.11e HCCA mode. *Wirel. Netw.* **2013**, *19*, 771–783.
75. de Haan, R.; Boucherie, R.J.; van Ommeren, J.-K. A polling model with an autonomous server. *Queueing Syst.* **2009**, *62*, 279–308.
76. van der Mei, R.D.; Roubos, A. Polling models with multi-phase gated service. *Ann. Oper. Res.* **2012**, *198*, 25–56.
77. van Wijk, A.C.C.; Adan, I.J.B.F.; Boxma, O.J.; Wierman, A. Fairness and efficiency for polling models with the k-gated service discipline. *Perform. Eval.* **2012**, *69*, 274–288.
78. Remerova, M.; Foss, S.; Zwart, B. Random fluid limit of an overloaded polling model. *Adv. Appl. Probab.* **2014**, *46*, 76–101.
79. Ling, Y.; Liu, C.; Li, Y. Study on queue strategy of gated polling multi-access communication system. *Recent Adv. Comput. Sci. Inf. Eng. Lect. Notes Electr. Eng.* **2012**, *124*, 99–105.
80. Vishnevskii, V.M.; Lakontsev, D.V.; Semenova, O.V.; Shpilev, S.A. A model of the polling system for studying the broadband wireless networks. *Autom. Remote Control* **2006**, *67*, 1974–1985.

81. Vatutin, V. A. Multitype Branching processes with immigration in random environment, and polling systems. *Sib. Adv. Math.* **2011**, *21*, 42–72.
82. Abidini, M.A.; Boxma, O.; Resing, J. Analysis and optimization of vacation and polling models with retrials. *Perform. Eval.* **2016**, *98*, 52–69.
83. Kim, B.; Kim, J. Analysis of the waiting time distribution for polling systems with retrials and glue periods. *Ann. Oper. Res.* **2019**, *277*, 197–212.
84. Dorsman, J.L.; van der Mei, R.D.; Winands, E.M.M. Polling systems with batch service. *OR Spectr.* **2012**, *34*, 743–761.
85. Boon, M.A.A.; Winands, E.M.M.; Adan, I.J.B.F.; van Wijk, A.C.C. Closed-form waiting time approximations for polling systems *Perform. Eval.* **2011**, *68*, 290–306.
86. Jiang, T.; Liu, L.; Zhu, Y. Analysis of a batch service polling system in a multi-phase random environment. *Methodol. Comput. Appl. Probab.* **2017**, 1–20, doi:10.1007/s11009-017-9585-0.
87. Shomrony, M.; Yechiali, U. *Polling Systems with Positive and Negative Customers*; Technical Report; Department of Statistics and Operations Research; Tel-Aviv University: Tel Aviv-Yafo, Israel, 2006.
88. Shomrony, M.; Yechiali, U. *Polling Systems with Job Failures and with Station Failures*; Technical Report; Department of Statistics and Operations Research; Tel-Aviv University: Tel Aviv-Yafo, Israel, 2006.
89. Zorine, A.V. On ergodicity conditions in a polling model with Markov modulated input and state-dependent routing. *Queueing Syst.* **2014**, *76*, 223–241.
90. Boon, M.A.A. A polling model with reneging at polling instants. *Ann. Oper. Res.* **2012**, *198*, 5–23.
91. Granville, K.; Drekic, S. On a 2-class polling model with reneging and k_i-limited service. *Ann. Oper. Res.* **2019**, *274*, 267–290.
92. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Johns Hopkins University Press: Baltimore, MD, USA, 1981.
93. Boxma, O.J.; Bruin, J.; Fralix, B.H. Sojourn times in polling systems with various service disciplines. *Perform. Eval.* **2009**, *66*, 621–639.
94. Bekker, R.; Vis, P.; Dorsman, J.L.; van der Mei, R.D.; Winands, E.M.M. The impact of scheduling policies on the waiting-time distributions in polling systems. *Queueing Syst. Theory Appl.* **2015**, *79*, 145–172.
95. Vis, P.; Bekker, R.; van der Mei, R.D. Heavy-traffic limits for polling models with exhaustive service and non-FCFS service order policies. *Adv. Appl. Probab.* **2015**, *47*, 989–1014.
96. Kim, B.; Kim, J. Sojourn time distribution in polling systems with processor-sharing policy. *Perform. Eval.n* **2017**, *114*, 97–112.
97. Cao, J.; Xie, W. Stability of a two-queue cyclic polling system with BMAPs under gated service and state-dependent time-limited service disciplines. *Queueing Syst.* **2016**, *85*, 117–147.
98. Chen, W.-L. Computing the moments of polling models with batch Poisson arrivals by transform inversion. *INFORMS J. Comput.* **2019**, *31*, 411–632.
99. Suman, R.; Krishnamurthy, A. Analysis of tandem polling queues with finite buffers. *Ann. Oper. Res.* **2019**, doi:10.1007/s10479-019-03358-0
100. Antunes, N.; Fricker, C.; Roberts, J. Stability of multi-server polling system with server limits. *Queueing Syst.* **2011**, *68*, 229–235.
101. Boxma, O.; van der Wal, J.; Yechiali, U. Polling with batch service. *Stoch. Model.* **2008**, *24*, 604–625.
102. Vlasiou, M.; Yechiali, U. $M/G/\infty$ polling systems with random visit times. *Probab. Eng. Infor. Sci.s* **2008**, *22*, 212–245.
103. van der Mei, R.D.; Winands, E.M.M. A note on polling models with renewal arrivals and nonzero switch-over times. *Oper. Res. Lett.* **2008**, *36*, 500–505.
104. van der Mei, R.D.; Levy, H. Polling systems in heavy traffic: Exhaustiveness of service policies. *Queueing Syst.* **1997**, *27*, 227–250.
105. Dorsman, J.L.; van der Mei, R.D.; Winands, E.M.M. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *Stoch. Model.* **2011**, *27*, 318–332.
106. Boon, M.A.A.; van der Mei, R.D.; Winands, E.M.M. Heavy traffic analysis of roving server networks. *Stoch. Model.* **2017**, *33*, 1–21.
107. Meyfroyt, T.M.M.; Boon, M.A.A.; Borst, S.C.; Boxma, O.J. Performance of large-scale polling systems with branching-type and limited service. *Perform. Eval.* **2019**, *133*, 1–24.
108. Kavitha, V.; Combes, R. Mixed polling with rerouting and applications. *Perform. Eval.* **2013**, *70*, 1001–1027.
109. Boxma, O.; Ivanovs, J.; Kosinski, K.; Mandjes, M. Levy-driven polling systems and continuous-state branching processes. *Stoch. Syst.* **2011**, *1*, 411–436.
110. Leskela, L.; Unger, F. Stability of a spatial polling system with greedy myopic service. *Ann. Oper. Res.* **2012**, *198*, 165–183.
111. Kavitha, V.; Altman, E. Continuous polling models and application to ferry assisted WLAN. *Ann. Oper. Res.* **2012**, *198*, 185–218.
112. Beekhuizen, P.; Denteneer, D.; Resing, J. Reduction of a polling network to a single node. *Queueing Syst.* **2008**, *58*, 303–319.
113. Matveev, A.; Feoktistova, V.; Bolshakova, K. On global near optimality of special periodic protocols for fluid polling systems with setups. *J. Optim. Theory Appl.* **2016**, *171*, 1055–1070.
114. Saffer, Z.; Telek, M.; Horvath, G. Fluid polling system with Markov modulated load and gated discipline. *Lect. Notes Comput. Sci.* **2018**, *10932*, 86–102.
115. Yechiali, U.; Czerniak, O. Fluid polling systems. *Queueing Syst.* **2009**, *63*, 401–435.
116. Czerniak, O.; Altman, E., Yechiali, U. Orchestrating parallel TCP connections: Cyclic and probabilistic polling policies. *Perform. Eval.* **2012**, *69*, 150–163.