Probability in the Engineering and Informational Sciences, **26**, 2012, 17–42.
doi:10.1017/S0269964811000210

WAIT-AND-SEE STRATEGIES IN POLLING MODELS

FRANK AURZADA, SERGEJ BECK, AND MICHAEL SCHEUTZOW
Technische Universität Berlin
Institut für Mathematik
10623 Berlin, Germany
E-mail: aurzada@math.tu-berlin.de; sergej.beck@hotmail.de;
ms@math.tu-berlin.de

We consider a general polling model with N stations. The stations are served exhaustively and in cyclic order. Once a station queue falls empty, the server does not immediately switch to the next station. Rather, it waits at the station for the possible arrival of new work ("wait-and-see") and, in the case of this happening, it restarts service in an exhaustive fashion. The total time the server waits idly is set to be a fixed, deterministic parameter for each station. Switchover times and service times are allowed to follow some general distribution, respectively. In some cases, which can be characterized, this strategy yields a strictly lower average queuing delay than for the exhaustive strategy, which corresponds to setting the "wait-and-see credit" equal to zero for all stations. This extends the results of Peköz [12] and of Boxma et al. [4]. Furthermore, we give a lower bound for the delay for *all* strategies that allow the server to wait at the stations even though no work is present.

1. INTRODUCTION AND MAIN RESULTS

1.1. Introduction

In this work, we consider a polling model in the sense of [13]. In a polling model, one server serves several queues, called stations. The classical service procedures are the (1) exhaustive, (2) gated, and (3) limited strategies, in which the server serves each station (1) until no more work is waiting at the respective station, (2) until all the work is served that was awaiting the server upon its arrival at the station, or (3) until the server has finished at most a predescribed number of jobs. The server then turns its attention to the next station. A possible (deterministic or random) idle time between the different stations, called switchover time, accounts for things like reloading or refueling.

Recently, a few articles (most importantly [4] and [12]; also see [3] and [16]) consider strategies in which the server does not immediately switch from one station to the next if the queue there is empty. Rather, it possibly waits at the station for a while for the potential arrival of new messages. This is particularly useful if (1) the switchover times are random with sufficiently large variances and (2) if the server is not likely to find much work at the other stations, i.e., if the traffic intensity of the current station is much larger than those of the other stations.

The strategy proposed and analysed in this article is also of this type: Each station i is given a fixed wait-and-see credit $T_i \geq 0$. Once the server arrives at station i, it will work there whenever messages are waiting, but it will also wait (and see) at the station for a total time of T_i. Once the credit is used up and no more messages are waiting, it will switch to the next station. This strategy was considered by Peköz [12] for the case of a completely symmetric system, for which arrival rates, service times, switchover times, and the T_i are identical for all stations.

The main contributions of this article are as follows:

- To extend Peköz' results to the general (i.e., not necessarily symmetric) polling model and to show, in particular, that the asymmetry induces some new effects previously not observed;
- To show that our strategy can be adjusted to provide lower delay than with the exhaustive strategy in several cases (which can be characterized and which also appear for deterministic switchover times);
- To analyze the case of a polling model with two stations in detail and compare our strategy to the one proposed by Boxma et al. [4];
- To prove a lower bound for the delay for *all* strategies that allow the server to wait at a station even though no work may be present.

As mentioned earlier, introducing a wait-and-see credit is particularly useful if the server is not likely to find much traffic at the other stations. This is because changing the station means to stay idle for a switchover time rather than resuming work at the current station within a short time. Surprisingly, we will see that this effect is largely independent of the length of the switchover times.

So far, the advantage of additional idle times—as we apply them here—was ascribed to the random switchover times. The new observation is that using nonzero idle times is also particularly useful if the system is asymmetric, that is, one of the stations experiences much more traffic than the others, even though the switchover times may be deterministic. This is an aspect that could not be observed in [12], and even though being intuitive, we can quantify this effect precisely.

In our polling model, the stations are served in cyclic order. We mention that the performance of all strategies can yet be improved by altering the order in which the server serves the different stations. For example, star polling can be applied if one of the stations experiences significantly more traffic than all other stations (see, e.g., [2,11]).

This article is structured as follows. In Section 1.2, we describe the model in detail and introduce the relevant parameters. Section 1.3 contains a summary of our main results. We review related work in Section 1.4. The proofs for the main results are given in Section 2 for the general case and in Section 3 for the refined results for polling models with only two stations. In Section 4, we prove a lower bound for the delay for *all* strategies that allow the server to wait at a station even though no work may be present. We highlight some possible further improvements and lines of future reseach in Section 5.

The motivation for this work comes from a real-world application. In so-called Ethernet Passive Optical Networks (EPONs; see [8,9]), a service provider is connected to various end users via an optical fiber cable. Different optical wavelength channels may be available on the cable for the communication, but each wavelength channel can be operated only either upstream (messages are sent from end users to the service provider) or downstream at a given time. Switching from upstream to downstream operation or vice versa incurs an idle time (switchover time). Therefore, each channel of an EPON can be regarded as a polling model.

1.2. The Model

We consider a polling model with $N \geq 1$ stations and one server that serves the stations in cyclic order. The stations are numbered $i = 1, \ldots, N$; because of the cyclic order, when we talk of the stations, we set $N + 1 \triangleq 1$.

Each station i has its own queue that is fed by a Poisson arrival process whose arrival rate is denoted by λ_i. Each arriving message has a random length (also called service time). The mean and second moment of the message length distribution are denoted by b_i and $b_i^{(2)}$, respectively, and are assumed to be finite.

The behavior of the server can be described as follows. The server arrives at station i and starts serving first come–first server [FCFS] all waiting messages and newly arriving messages until the queue is empty. This is typically called exhaustive service in the context of polling models. However, once the station is empty or if the server finds an empty station upon its arrival, the server does not immediately switch to the next station; rather it turns idle for some time in order to wait for potentially newly arriving messages ("wait-and-see"). As soon as new messages arrive, it starts serving them immediately and in an exhaustive fashion. Once finished, it again turns idle and waits for new messages to arrive, and so on.

The main feature of our model is that the server is set to wait idly for new messages for a *total* time T_i, where $T_i \geq 0$ is a fixed parameter of the system, called wait-and-see credit. This total time can be spent altogether in one single period (e.g., if there are no messages waiting at the station upon the server's arrival and no messages arrive even until time T_i after the server's arrival at the station) or it can be spent in different periods—interleaved by different busy periods. Note that since T_i is fixed, the server might not use any information about the current queue status at other stations or about the future of the arrival process at any station.

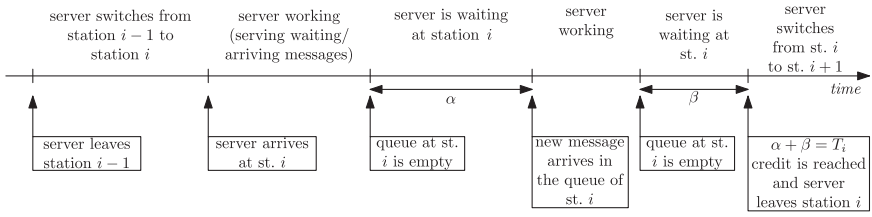

FIGURE 1. Operation of our polling model.

After the server has spent a total waiting time of T_i at station i, it starts the switchover to station $i+1$. Hereby, it first spends a possibly random idle time, called switchover time, where it does not serve any messages neither at station i or at station $i+1$. The random switchover time from station i to station $i+1$ is assumed to have finite mean $r_i \geq 0$ and finite second moment $r_i^{(2)}$. We will consider both nondeterministic and deterministic switchover times (in the latter case, $r_i^{(2)} = r_i^2$).

The message generation process, the lengths of the messages, and the switchover times are assumed to be independent—both among each other and with respect to the other processes and stations. An illustration of the operation of the server is given in Figure 1.

The goal of this article is to derive an explicit formula for the mean average delay of a job for this model in steady state—that is, the expected time a message experiences from the point in time when it arrives in one of the queues until its service starts (i.e., excluding the processing time). The expected delay of a message generated at station i is denoted by $\mathbb{E}D_i$; the mean average queuing delay is then defined by

$$\bar{D} = \sum_{i=1}^{N} \frac{\rho_i}{\rho_0} \mathbb{E}D_i,$$

where $\rho_i := \lambda_i b_i$ is the traffic load offered to station i and $\rho_0 := \sum_{i=1}^{N} \rho_i$ is the total load offered to the system. We stress that the delays of the different stations are weighted by the traffic intensity ρ_i, which implicitly includes weighting with the average message lengths, whereas the delays $\mathbb{E}D_i$ do not include weighting the delay of the individual messages with their lengths. This seems to be common in the literature; for example Takagi [13, p. 92] calls this quantity *intensity weighted mean waiting time*.

Note that under weighting by the intensities ρ_i that we apply here, it is clear that finishing work at the present station (exhaustively) is necessarily better than leaving work undone at the station. This observation is not necessarily correct if one uses weighting e.g. by the arrival rates λ_i, since then it might be more efficient to abandon work at a station and to switch to another station where presumably many small messages are waiting (i.e., where λ_i is large, even though ρ_i might be small).

The mean delay will be expressed explicitly in terms of the parameters λ_i, b_i, $b_i^{(2)}$, T_i, r_i, and $r_i^{(2)}$, $i = 1, \ldots, N$, defined earlier.

Furthermore, it will be convenient to use the following abbreviations. We abbreviate by $r_0 := \sum_{i=1}^N r_i$ the sum of the mean switchover times and by $r_0^{(2)} := \sum_{i=1}^N r_i^{(2)} + \sum_{i,j=1, i\neq j}^N r_i r_j$ the second moment of the sum of all switchover times. Finally, we let $T_0 := \sum_{i=1}^N T_i$ denote the total "wait-and-see" time per cycle.

1.3. Main Results

In this subsection, we give our main results. Theorem 1 gives a formula for the mean average delay in terms of the parameters of the system λ_i, b_i, $b_i^{(2)}$, r_i, and $r_i^{(2)}$, $i = 1, \ldots, N$, as well as the times T_i, $i = 1, \ldots, N$. This is simplified for the case of two stations, $N = 2$, in Corollary 2.

The formula for the delay allows one to investigate the following question: Given the system parameters λ_i, b_i, $b_i^{(2)}$, r_i, and $r_i^{(2)}$, how does one have to adjust the parameters T_i, $i = 1, \ldots, N$, such that the mean average delay is minimized. We will see that in many cases it is favorable—in the sense of lower average queueing delay – to choose *positive* T_i. This is either due to (1) the random switchover times or (2) the asymmetry of the system. This is described in detail for the case of two stations in Theorems 3 and 4, in which effects (1) and (2) are treated somehow in a decoupled way.

Finally, in Theorem 8 we consider all strategies that allow the server to wait at a station even though no work might be present. We give a lower bound for the delay for any such strategy.

Before we come to the main results, let us mention the stability condition for the system. Due to the exhaustive nature of our polling model, it is clear that the system is stable if and only if

$$\rho_0 < 1.$$

We assume this condition from now on.

Furthermore, we recall the result for the so called exhaustive strategy from [13]. In our model, this corresponds to the special case $T_1 = T_2 = \cdots = T_N = 0$. In this case, one can find that

$$\bar{D} = \frac{\sum_{i=1}^N \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{r_0(\rho_0^2 - \sum_{i=1}^N \rho_i^2)}{2\rho_0(1-\rho_0)} + \frac{r_0^{(2)}}{2r_0}.$$

In terms of queuing delay, it was shown in [10] that the exhaustive strategy provides the lowest delay in the class of all nonidle strategies, in particular, as compared to the gated and limited strategy. Therefore, it will serve as a benchmark for our strategy, which is a strategy allowing the server to be idle even though work might be present in the system (at other stations).

The main theorem for our polling model is as follows.

THEOREM 1: *The mean average delay of the polling model introduced above is given by*

$$\bar{D} = \frac{\sum_{i=1}^N \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{(r_0+T_0)(\rho_0^2 - \sum_{i=1}^N \rho_i^2)}{2\rho_0(1-\rho_0)} + \frac{\frac{1}{2}\rho_0 r_0^{(2)} + r_0 \sum_{i=1}^N T_i(\rho_0-\rho_i)}{\rho_0(r_0+T_0)} + \frac{1}{(r_0+T_0)\rho_0}\left[\sum_{i=1}^N T_i^2 \frac{(1-2\rho_i)(\rho_0-\rho_i)}{2(1-\rho_i)} + \sum_{1\le i<j\le N} T_i T_j(\rho_0-\rho_i-\rho_j)\right]. \tag{1}$$

The proof of this theorem is given in Section 2. Certainly, one can ask which values of $T_1, \dots, T_N$ lead to a minimal queueing delay. In other words, given the system parameters, we would like to know how we have to set $T_1, \dots, T_N$ in order to minimize $\bar{D}$. Note that this is a nontrivial question, because the T_i appear in the numerator and denominator. In fact, this is a minimization problem in the variables $T_1, \dots, T_N$, subject to the nonnegativity restriction $T_i \ge 0$, for all $i = 1, \dots, N$, which can be carried out—in principle—explicitly. We discuss the respective minimizers below for $N = 2$. Certainly, for large N, one would solve the problem numerically.

For two stations $N = 2$, the main result reduces to the following simpler formula.

COROLLARY 2: *The mean average delay of the polling model introduced above with $N = 2$ is given by*

$$\bar{D} = \frac{\sum_{i=1}^2 \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{1}{\rho_0(r_0+T_0)}\left[\frac{r_0^{(2)}\rho_0}{2} + \frac{\rho_1\rho_2}{1-\rho_0}(r_0+T_0)^2 + \rho_2 T_1\left(r_0 + T_1\frac{1-2\rho_1}{2(1-\rho_1)}\right) + \rho_1 T_2\left(r_0 + T_2\frac{1-2\rho_2}{2(1-\rho_2)}\right)\right]. \tag{2}$$

In particular, one can minimize (2) with respect to T_1 and T_2 subject to the restrictions $T_1 \ge 0$ and $T_2 \ge 0$ in order to obtain the minimal possible delay. Let us denote by T_1^* and T_2^* the minimizers. We say that there is *no gain from waiting at station i* if $T_i^* = 0$; if $T_i^* > 0$, we say that it is *worth waiting at station i*.

From the above, explicit expression one can observe the following consequences. First, we consider a partially symmetric polling model, by which we *only* mean that both stations have the same intensities $\rho_1 = \rho_2$ but not necessarily the same switchover time distribution, message length distribution, or arrival rate (cf. [12]).

THEOREM 3: *Consider a polling model as introduced above with two stations. In the case of a symmetric polling model, $\rho_1 = \rho_2$, the following holds:*

- *With deterministic switchover times (i.e., $r_1^2 = r_1^{(2)}$ and $r_2^2 = r_2^{(2)}$), we get $T_1^* = T_2^* = 0$; that is, in this case, there is no gain from waiting at either station.*

- *With nondeterministic switchover times (i.e., $r_1^2 < r_1^{(2)}$ or $r_2^2 < r_2^{(2)}$), it is worth waiting (at both stations) if and only if*

$$2\rho_1 < 1 - \frac{r_0^2}{r_0^{(2)} + r_0^2\rho_1/(1-2\rho_1)}. \tag{3}$$

In this case, the optimal waiting time $T_1^* = T_2^* > 0$ can be calculated explicitly, see (27). The minimal delay is then given by inserting (27) into (2); this delay is strictly lower than the mean average delay induced by the exhaustive strategy.

Now, we consider an asymmetric polling model (i.e., $\rho_1 > \rho_2$).

THEOREM 4: *Consider a polling model as introduced above with two stations. In the case of an asymmetric polling model with deterministic switchover times (i.e., assuming $\rho_1 > \rho_2$ and $r_1^2 = r_1^{(2)}$, $r_2^2 = r_2^{(2)}$), the following holds:*

- *There is no gain from waiting at station 2 (i.e., in all cases $T_2^* = 0$).*
- *Further, it is worth waiting at station 1 if and only if*

$$\rho_1 - \rho_1^2 + \rho_2^2 - \rho_2 - 2\rho_1\rho_2 > 0. \tag{4}$$

In this case, one can calculate the minimizer $T_1^* > 0$ explicitly as in (30). The minimal delay is then given by inserting (30) and $T_2^* = 0$ into (2); this delay is strictly lower than the mean average delay induced by the exhaustive strategy.

Remark 5: A similar discussion is possible for the case of an asymmetric polling model with nondeterministic switchover times. There, both of the following effects will be combined; namely note that Theorem 3 shows that large variances of the switchover times lead to the situation in which it is worth waiting (at both stations): Indeed, the fraction on the right-hand side of (3) equals

$$\frac{(\mathbb{E}[R_1 + R_2])^2}{\mathrm{var}[R_1 + R_2] + (\mathbb{E}[R_1 + R_2])^2(1-\rho_1)/(1-2\rho_1)}, \tag{5}$$

where R_i are independent switchover times for switching from station i to station $i+1$, respectively. Thus, increasing the variances of the switchover times in condition (3) makes the condition less strict; thus, it becomes worth waiting.

On the other hand, Theorem 4 shows that a strong asymmetry (reducing ρ_2 in condition (4)) makes it useful to wait at the station with significantly higher traffic intensity.

These effects will both be present in the case of a not necessarily symmetric system with nondeterministic switchover times.

Remark 6: We remark the following rather surprising fact: Note that (4) does not depend on the switchover times. So, the question of whether it makes sense to wait

at station 1 only depends on the relation of the intensities ρ_i, not on the length of the possible idle period due to the switching. Similarly, the expression in (3) does not depend on the absolute lengths of the two switchover times (one can multiply both R_i by the same constant without changing (3), cf. (5)) or on the order of the switchover times (but only on the sum). However, even though the decision whether to wait or not *does not* depend on the absolute length of the switchover time, the resulting credit (i.e., the optimal waiting time) *does* (cf. (27)).

Remark 7: If $N > 2$, the mean average queueing delay $\bar{D}$ has the form

$$\bar{D} = c + \frac{\vec{T}A\vec{T}^t + \vec{T}\vec{b}^t + a}{r_0 + T_0}$$

with some $N \times N$ matrix A, constants c, a, and $\vec{b}$, and $\vec{T} := (T_1, \ldots, T_N)$. One can again discuss for which values of $\vec{T}$ the delay is minimized and find the minimizers. Note that from the point of view of applications, the above form allows for an easy numerical determination of the minimal delay and the respecitive minimizer $\vec{T}^*$.

Finally, we discuss a lower bound for the delay for strategies that allow waiting times of any type. We recall at this point that [10] shows that the exhaustive strategy provides the lowest delay in the class of all nonidle strategies—that is, all strategies in which the server is *not* allowed to wait at a station if no work is present there.

In the following, we consider strategies that are not allowed to use future information of the system, that serve FCFS, and where the server is not idle if at its present station messages are waiting to be served. Further, we have to assume that with this strategy the system has a steady-state distribution.

The next theorem gives a lower bound for the delay for all of these strategies for which the server *is* allowed to wait at stations due to reasons that depend only on the current station in the current cycle (i.e., since the server arrived at the present station). This restriction considers those strategies that look at the evolution of the traffic at the present station since the server arrived there. It does not allow strategies that take their decisions according to, for example, the queue status at different stations or the recent switchover times.

This provides a lower bound, in particular, for the model treated so far, the strategy proposed in [4], as well as the strategy proposed in Section 5.

THEOREM 8: *Consider a polling model for which the stations are served in cyclic order. Then for any strategy that allows the server to wait at a station even though no work is present there but the decision on whether and how long to wait only depends on the evolution of queue of the current station since the server arrived at the station,*

we have

$$\bar{D} \geq \frac{\sum_{i=1}^{N} \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{r_0(\rho_0^2 - \sum_{i=1}^{N} \rho_i^2)}{2\rho_0(1-\rho_0)} + \min_{f_1,\ldots,f_N \geq 0} \frac{1}{\rho_0(r_0+f_0)} \left[\frac{\rho_0 r_0^{(2)}}{2} + \sum_{i=1}^{N} \left(r_0 f_i + \frac{f_i^2}{2} \right) (\rho_0 - \rho_i) + \sum_{i=1}^{N} f_i \left(\sum_{j=1}^{i-1} \alpha_j \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j-1} \rho_l \right) + \sum_{j=i+1}^{N} \alpha_j \sum_{l=i+1}^{j-1} \rho_l \right) \right], \tag{6}$$

where $\alpha_j := \rho_j(r_0+f_0)/(1-\rho_0) + f_j$ *and* $f_0 = \sum_{i=1}^{N} f_i$.

The idea behind this theorem is that f_i is the expected time the server spends waiting at station i in a cycle. The time as such is random for a general strategy, of course. Since the f_i are unknown in general, the minimum appears. In the case of the concrete model treated so far, we had $f_i = T_i$, because by definition the total time the server spends at station i is deterministic and equals T_i.

The minimum in (6) can be calculated explicitly as well as numerically without any problem. We note that the term in (6) that is to be minimized actually has the same format as the term in (1), with the f_i instead of T_i. The general format of this term is displayed in Remark 7; for $N = 2$, the discussions to obtain the minimum are completely analogous to those that lead to Theorems 3 and 4. The proof of Theorem 8 is given in Section 4.

1.4. Related Work

Basic references on polling models are [10,13–15].

References that refer to polling models in which the server might be waiting at a station are apparently rare. The main references for us are Peköz [12] and Boxma et al. [4].

Peköz [12] introduced the strategy we use in this article for the completely symmetric model (i.e., all of the arrival rates, service times, switchover times, and the T_i are identical). In particular, his Theorem 2.2 is a special case of our Theorem 1. In the present article, we consider the general polling model. Furthermore, a new observation is that also a sufficiently asymmetric system can make it useful to wait at a station, independently of whether the switchover times are random.

The second main reference is Boxma et al. [4], in which a polling model with $N = 2$ stations is analyzed. In that work, the following situation is investigated. If the server encounters an *empty queue* at station 1 *once it arrives there*, a "wait-and-see" timer is activated in order to wait for the possible arrival of new messages. However—contrary to the present setup—once the server has cleared the station of waiting messages (exhaustively) or the timer has run out, it will immediately switch

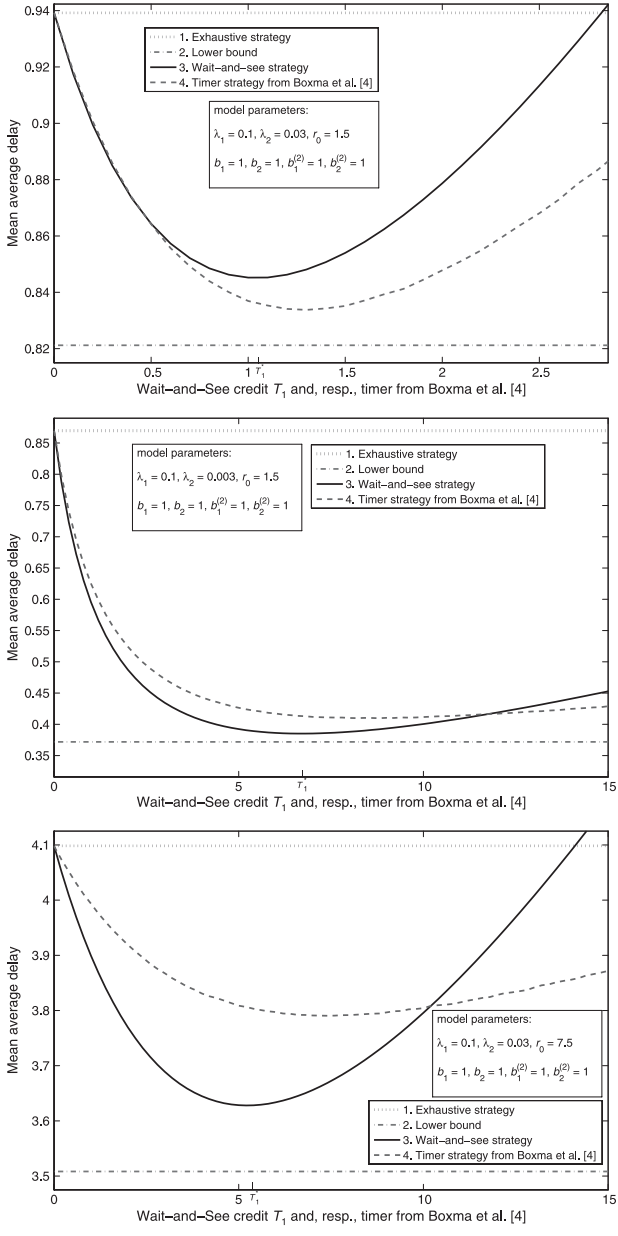

FIGURE 2. Comparison to Boxma et al. [4], delay versus credit (resp. timer) at station 1. Note that in the first plot, the strategy from [4] provides lower delay; in the second plot, only the arrival rate of the second station is changed and then our wait-and-see strategy has lower delay; in the third plot, the same happens due to longer switchover times.

to the next station. We compare the resulting delay obtained from this strategy to ours in Figure 2. We have found cases in which our strategy leads to lower delay than the strategy proposed by Boxma et al. and also cases in which it performs worse. The latter is usually the case if the intensities ρ_1 and ρ_2 are close to each other, whereas in the case that we deal with a highly asymmetric system, our strategy seems to be better. Additional, for large switchover times, our strategy seems to perform better than [4], since in this case the timer from [4] is rarely activated. Unfortunately, it does not seem to be possible to compare the strategies directly due to the nonexplicit nature of the delay formulas in [4].

Further references on polling models in which the server might be waiting, consider only single-station systems with vacations ([3,16]).

The strategy employed in [4] and in the present article is somehow related to a so-called *forced idle time*. We refer, for example, to [5,6] for some work on this. However, in the present setup, the server is not forced to be idle; whenever it is set to "wait-and-see," it resumes service as soon as new messages arrive. This is the reason we prefer the term "wait-and-see" rather than "forced idle time."

2. THE FUNDAMENTAL RELATIONS FOR THE GENERAL POLLING MODEL

In this section, we derive the fundamental relations for the general polling model that allow us to obtain the formula for the mean average queueing delay. In particular, we give a proof of Theorem 1. We proceed in several steps.

Cycle Time

The cycle time is defined to be the time that the server takes from one arrival at station 1 to its next arrival at the same station. We obtain the average cycle time $\mathbb{E}C$.

First, let us define more precisely the notion of the server being *idle*, *switching*, and *waiting*. The server is *waiting* when it is at some station waiting for messages to arrive. Note that, by the definition of our strategy, the total time the server spends waiting in each cycle equals $\sum_{i=1}^{N} T_i$. The server is said to be in the state of *switching* from the time it leaves one station until it starts working at the next station. Finally, we say that the server is *idle* if it is either waiting or switching.

Now, note that a cycle contains periods when the server works and periods when it is idle. In our polling model, the server is idle exactly for the time it waits and for the switchover time. Thus, the expected time the server is idle in a cycle equals

$$(1-\rho_0)\mathbb{E}C = \sum_{i=1}^{N} r_i + \sum_{i=1}^{N} T_i. \tag{7}$$

This allows one to deduce the expected cycle time in our polling model in steady state:

$$\mathbb{E}C = \frac{r_0 + T_0}{1-\rho_0}. \tag{8}$$

The Decomposition Principle

We will use a decomposition principle to deduce our results. One can find a proof of this principle in other contexts (e.g., in [1,4]). We omit the proof for our system since it is completely analogous.

In order to formulate the decomposition principle, we need the notion of the *workload* V, which we define to be the sum of all message lengths that are present in the system (including the remaining length of the currently processed message) at a random point in time in steady state.

Then the decomposition principle says that V has the same distribution as

$$V \stackrel{d}{=} V^{M/G/1} + V^{\text{idle}},$$

where $V^{M/G/1}$ is the workload in the same polling model with no switching or waiting times, (i.e., an $M/G/1$ queue). On the other hand, V^{idle} is the workload at a random point in time given that the server is idle at that point, and $V^{M/G/1}$ and V^{idle} are independent.

Let

$$q := \mathbb{P}\left(\text{server switching} \mid \text{server idle}\right) = \frac{\mathbb{P}\left(\text{server switching}\right)}{\mathbb{P}\left(\text{server switching}\right) + \mathbb{P}\left(\text{server waiting}\right)}.$$

Therefore,

$$\mathbb{E}V = \mathbb{E}V^{M/G/1} + q\mathbb{E}V^{\text{switching}} + (1-q)\mathbb{E}V^{\text{waiting}}, \tag{9}$$

where $V^{\text{switching}}$ and V^{waiting} are the workloads at a random point in time given that the server is switching and waiting, respectively, at that point.

Expected Workload

We now calculate $\mathbb{E}V$ in two different ways. On the one hand, note that

$$\mathbb{E}V = \sum_{i=1}^{N} b_i \mathbb{E}[\#\text{ messages in queue at station } i] + \sum_{i=1}^{N} \rho_i \frac{b_i^{(2)}}{2b_i}. \tag{10}$$

Indeed, the first term accounts for the fact that there are messages that are not yet in service and that are waiting at the different stations. The second term corresponds to the fact that with probability ρ_i we are looking at station i and a message is being processed there. The workload of that message is exactly its expected residual lifetime, i.e., $b_i^{(2)}/(2b_i)$.

By Little's law, (10) becomes

$$\mathbb{E}V = \sum_{i=1}^{N} b_i \lambda_i \mathbb{E}D_i + \sum_{i=1}^{N} \rho_i \frac{b_i^{(2)}}{2b_i} = \sum_{i=1}^{N} \rho_i \mathbb{E}D_i + \sum_{i=1}^{N} \rho_i \frac{b_i^{(2)}}{2b_i} = \rho_0 \bar{D} + \sum_{i=1}^{N} \rho_i \frac{b_i^{(2)}}{2b_i}. \tag{11}$$

This equation shows that in order to obtain the mean average queuing delay of our system, $\bar{D}$, we have to calculate the expected workload $\mathbb{E}V$.

On the other hand, we use the decomposition principle (9). Clearly, $\mathbb{E}V^{M/G/1}$ is known:

$$\mathbb{E}V^{M/G/1} = \frac{\sum_{i=1}^{N} \lambda_i b_i^{(2)}}{2(1-\rho_0)};$$

see, for example, [7, p. 201]. Therefore, we obtain with (9) and (11) that

$$\rho_0 \bar{D} = \frac{\sum_{i=1}^{N} \lambda_i b_i^{(2)}}{2(1-\rho_0)} + q\mathbb{E}V^{\text{switching}} + (1-q)\mathbb{E}V^{\text{waiting}} - \sum_{i=1}^{N} \rho_i \frac{b_i^{(2)}}{2b_i}. \tag{12}$$

We are left with calculating the expected workload given that we find the system in the state of switching and, respectively, the expected workload given that we find the system in the state of waiting, as well as some relation between time periods of switching and waiting. Concerning the latter, in fact, it is sufficient that we clearly know that

$$\frac{q}{\mathbb{P}\,(\text{server switching})} = \frac{1}{\mathbb{P}\,(\text{server switching}) + \mathbb{P}\,(\text{server waiting})} = \frac{1}{1-\rho_0} \tag{13}$$

and

$$\frac{1-q}{\mathbb{P}\,(\text{server waiting})} = \frac{1}{\mathbb{P}\,(\text{server switching}) + \mathbb{P}\,(\text{server waiting})} = \frac{1}{1-\rho_0}. \tag{14}$$

Workload Present While Switching

Observe that

$$\mathbb{E}V^{\text{switching}} = \frac{1}{\mathbb{P}\,(\text{server switching})} \sum_{i=1}^{N} p_i \mathbb{E}V_i^{\text{switching}}, \tag{15}$$

where p_i is the probability of encountering the server in the state of switching from station i to station $i+1$ when entering the system at a random point in time and $\mathbb{E}V_i^{\text{switching}}$ is the expected workload at such a point in time.

Clearly, $p_i = r_i/\mathbb{E}C$, since this is the fraction of time in a cycle that the server spends switching from station i to station $i+1$.

Now, we have to find $\mathbb{E}V_i^{\text{switching}}$, the expected total amount of work that is present given that we look at the system at a point when the server is switching from station i to station $i+1$. Let us assume that we are at such a point in time; then there are different times when the currently present workload was generated. We distingish these times and determine the respective workload:

- At all stations j, workload was generated during the current switching period. It is given by the expected residual lifetime of the current switching period: $r_i^{(2)}/(2r_i)\sum_{j=1}^{N} \rho_j$. Here, $r_i^{(2)}/(2r_i)$ is the expected residual lifetime (in fact, the expected backward recurrence time) of the current (ith) switching period; this becomes weighted by $\sum_{j=1}^{N} \rho_j$, the sum of the intensities at all stations.

- At all stations j, except for the ith station, workload was generated while the server was at station i (working and waiting). The time spent working has mean $\rho_i \mathbb{E}C$ and the time spent waiting equals T_i, so that the generated workload becomes $(\rho_i \mathbb{E}C + T_i) \sum_{j=1,\, j\neq i}^{N} \rho_j$.
- Similarly, while the server was at station k (working and waiting), $k \neq i$, at all stations j (except for those that later in the cycle became emptied; i.e. only those stations that are situated "between station i and station k" in the cycle have to be considered) workload was generated and is still present. The time spent working has mean $\rho_k \mathbb{E}C$ and the time spent waiting equals T_k, so that the generated workload becomes

$$(\rho_k \mathbb{E}C + T_k) \sum_{j \in \{i+1,\ldots,k-1\}} \rho_j, \tag{16}$$

  where $\{i+1,\ldots,k-1\}$ is defined as $\{i+1,\ldots,N\} \cup \{1,\ldots,k-1\}$ if $i+1 > k-1$.
- During the switchover time from station k to station $k+1$ ($k \neq i$), which takes on average r_k, workload was generated at all stations (except for those that later in the cycle became emptied):

$$r_k \sum_{j \in \{i+1,\ldots,k\}} \rho_j. \tag{17}$$

Summing up all this workload, we get

$$\begin{aligned}
\mathbb{E}V_i^{\text{switching}} = {} & \sum_{j<i} r_j \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j} \rho_l \right) + \sum_{j>i} r_j \sum_{l=i+1}^{j} \rho_l \\
& + \sum_{j<i} \rho_j \mathbb{E}C \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j-1} \rho_l \right) + \sum_{j>i} \rho_j \mathbb{E}C \sum_{l=i+1}^{j-1} \rho_l \\
& + \sum_{j<i} T_j \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j-1} \rho_l \right) + \sum_{j>i} T_j \sum_{l=i+1}^{j-1} \rho_l \\
& + \rho_i \mathbb{E}C(\rho_0 - \rho_i) + (\rho_0 - \rho_i)\, T_i + \rho_0 \frac{r_i^{(2)}}{2r_i}.
\end{aligned} \tag{18}$$

Workload Present While Waiting

Analogous to the workload while switching, we observe that

$$\mathbb{E}V^{\text{waiting}} = \frac{1}{\mathbb{P}\,(\text{server waiting})} \sum_{i=1}^{N} q_i \mathbb{E}V_i^{\text{waiting}}, \tag{19}$$

where q_i is the probability of finding the server waiting (in a wait-and-see state) for messages at station i and $\mathbb{E}V_i^{\text{waiting}}$ is the workload one would find at such a point in time.

Clearly, $q_i = T_i/\mathbb{E}C$, since this is the fraction of time in a cycle that the server spends waiting at station i, by the definition of our polling model.

Similarly to the workload while switching, we obtain the expected workload generated while station i is in the state of waiting. Let us assume we are at such a point in time; then there are different times when the currently present workload was generated. We distinguish these times and determine the respective workload:

- When the server started working at station i, there was work waiting there. We denote the length of the "busy period" generated by this waiting traffic by Z_i. The workload generated at the other stations j, $j \neq i$, during this busy period is

$$\mathbb{E}Z_i \sum_{j \neq i} \rho_j = \mathbb{E}Z_i(\rho_0 - \rho_i). \tag{20}$$

  In order to determine $\mathbb{E}Z_i$ note that the average time the server spends working at station i is, on the one hand, $\rho_i \mathbb{E}C$. On the other hand, the time the server spends working at station i consists of the length of the first busy period, $\mathbb{E}Z_i$, and all other busy periods generated, which is the number of busy periods in an $M/G/1$ queue with total idle time T_i. However, the expected number of busy periods in an $M/G/1$ queue with total idle time T_i is $\lambda_i T_i$ (just disregard the time of the busy periods, then because of the memoryless property, the number of busy periods is Poisson with rate $\lambda_i T_i$).

  Thus,

$$\rho_i \mathbb{E}C = \mathbb{E}Z_i + \lambda_i T_i \frac{b_i}{1 - \rho_i}$$

  since $b_i/(1 - \rho_i)$ is the average length of one busy period at station i. This equation allows one to calculate $\mathbb{E}Z_i$ and to insert it into (20).

- Workload was generated at all other stations except for the ith, during the busy periods that have already taken place at station i, not considering the first busy period when the server started to work at station i (this was considered in the above point). Per busy period, a workload at the other stations j, $j \neq i$, of in total

$$\frac{b_i}{1 - \rho_i} \sum_{j \neq i} \rho_j = \frac{b_i}{1 - \rho_i}(\rho_0 - \rho_i)$$

  was generated, since $b_i/(1 - \rho_i)$ is the average length of one busy period at station i. In order to obtain the number of busy periods that have already taken place, note that these are, on average, $\lambda_i T_i/2$, because the waiting time

is deterministic. Thus, we obtain

$$\frac{\lambda_i T_i}{2}\frac{b_i}{1-\rho_i}\sum_{j\neq i}\rho_j = \frac{\lambda_i T_i}{2}\frac{b_i}{1-\rho_i}(\rho_0-\rho_i) \tag{21}$$

for the total workload that was generated at all other stations during the busy periods (except for the very first one) at station i.

- The total workload that was generated at all other stations during the waiting time spent so far at station i (recall that this is, on average, $T_i/2$, since it is deterministic) is

$$\frac{T_i}{2}\sum_{j\neq i}\rho_j = \frac{T_i}{2}(\rho_0-\rho_i). \tag{22}$$

- While the server was at station k (working and waiting), $k \neq i$, at all stations j (except for those that later in the cycle became emptied; i.e. only those stations that are situated "between station i and station k" in the cycle have to be considered) workload was generated and is still present. Thus the term in (16) has to be considered in the same way.
- During the switchover time from station k to station $k+1$ ($k \neq i$), the workload was generated at all stations (except for those that later in the cycle became emptied). Thus, the term in (17) has to be considered in the same way.

Summing up all of this workload, we get

$$\begin{aligned}\mathbb{E}V_i^{\text{waiting}} &= \sum_{j<i} r_j\left(\sum_{l=i+1}^{N}\rho_l + \sum_{l=1}^{j}\rho_l\right) + \sum_{j>i} r_j \sum_{l=i+1}^{j}\rho_l \\ &\quad + \sum_{j<i}\rho_j\mathbb{E}C\left(\sum_{l=i+1}^{N}\rho_l + \sum_{l=1}^{j-1}\rho_l\right) + \sum_{j>i}\rho_j\mathbb{E}C\sum_{l=i+1}^{j-1}\rho_l \\ &\quad + \sum_{j<i} T_j\left(\sum_{l=i+1}^{N}\rho_l + \sum_{l=1}^{j-1}\rho_l\right) + \sum_{j>i} T_j\sum_{l=i+1}^{j-1}\rho_l \\ &\quad + (\rho_0-\rho_i)\left(\rho_i\mathbb{E}C + \frac{T_i}{2}\left(1-\frac{\rho_i}{1-\rho_i}\right)\right).\end{aligned} \tag{23}$$

PROOF OF THEOREM 1: In order to see the formula in Theorem 1, one just has to combine (12), (8), and (13), (15), (18), and (14), (19), (23). ■

3. THE CASE OF TWO STATIONS

In this section, we prove Theorems 3 and 4. First, we prove that the optimal parameters in the two-station case satisfy a linear relation. Then we prove Theorem 3 (symmetric case, deterministic and nondeterministic) and Theorem 4 (asymmetric and deterministic).

For simplicity, we introduce the following abbreviations:

$$
\begin{aligned}
c_1 &:= \frac{\sum_{i=1}^{2} \lambda_i b_i^{(2)}}{2(1-\rho_0)},\\
c_2 &:= \frac{\rho_1\rho_2 r_0^2}{1-\rho_0} + \frac{\rho_0 r_0^{(2)}}{2},\\
c_3 &:= r_0\rho_2 + \frac{2\rho_2\rho_1 r_0}{1-\rho_0},\\
c_4 &:= r_0\rho_1 + \frac{2\rho_1\rho_2 r_0}{1-\rho_0},\\
c_5 &:= \frac{2\rho_2\rho_1}{1-\rho_0},\\
c_6 &:= \frac{c_5}{2} + \frac{\rho_2}{2}\left(1 - \frac{\rho_1}{1-\rho_1}\right),\\
c_7 &:= \frac{c_5}{2} + \frac{\rho_1}{2}\left(1 - \frac{\rho_2}{1-\rho_2}\right).
\end{aligned}
$$

An easy calculation shows that these are nonnegative constants. With these abbreviations, (2) becomes

$$
\bar{D} = c_1 + \frac{c_2 + c_3T_1 + c_4T_2 + c_5T_1T_2 + c_6T_1^2 + c_7T_2^2}{\rho_0(r_0 + T_1 + T_2)}. \tag{24}
$$

LEMMA 9: *On the set $r_0 + T_0 > 0$, the minimizers of the quantity in (24) satisfy the following linear relation:*

$$
(c_5 - 2c_6)T_1^* = c_3 - c_4 + (c_5 - 2c_7)T_2^*. \tag{25}
$$

In particular, in the symmetric polling model ($\rho_1 = \rho_2$), we must have

$$
T_1^* = T_2^*. \tag{26}
$$

We remark that the above minimizers can be negative. Recall that we are interested in the optimal parameter, which are the minimizers of (24) subject to the restriction $T_1^*, T_2^* \geq 0$. This is why we distinguish in the following between the minimizers of (24) and the optimal parameters.

PROOF OF LEMMA 9: Clearly, (24) shows that $\bar{D}$ can be written as

$$\bar{D} = \bar{D}(T_1, T_2) = c_1 + \frac{f(T_1, T_2)}{\rho_0(r_0 + T_1 + T_2)}$$

with some function f. If $\bar{D}$ has a minimum at T_1^* and T_2^* (with $r_0 + T_1^* + T_2^* > 0$), it must satisfy

$$\frac{\partial \bar{D}}{\partial T_1}(T_1^*, T_2^*) = 0 \quad \text{and} \quad \frac{\partial \bar{D}}{\partial T_2}(T_1^*, T_2^*) = 0.$$

Due to the fact that the denominator is a linear function in $T_1 + T_2$, an easy calculation shows that we must actually have

$$\frac{\partial f}{\partial T_1}(T_1^*, T_2^*) = \frac{\partial f}{\partial T_2}(T_1^*, T_2^*).$$

This is

$$c_3 + c_5 T_2^* + 2c_6 T_1^* = c_4 + c_5 T_1^* + 2c_7 T_2^*,$$

exactly as asserted in (25).

In the symmetric case, we have $\rho_1 = \rho_2 < 1/2$ and $c_3 = c_4$ and $c_7 = c_6$, which implies that (25) becomes (26). ■

3.1. Symmetric Polling Model

We now consider a symmetric polling model (i.e., $\rho_1 = \rho_2 =: \rho$).

PROOF OF THEOREM 3: Assume that $T_1^* > 0$ and $T_2^* > 0$ are the optimal parameters. Then we know from (26) that $T_1^* = T_2^* =: T$. Therefore, we obtain

$$\bar{D} = c_1 + \frac{\begin{array}{c}(r_0^{(2)} + \rho r_0^2/(1-2\rho)) + 2(r_0 + 2\rho r_0/(1-2\rho))T \\ +(4\rho/(1-2\rho) + 1 - \rho/(1-\rho))T^2\end{array}}{2(r_0 + 2T)}.$$

The minimum of this expression is attained at

$$T^* = -\frac{1}{2} r_0 + \frac{1}{2}\sqrt{4r_0^2\rho - 3r_0^2 + \left(r_0^{(2)} + r_0^2 \frac{\rho}{1-2\rho}\right)\left(4 - 12\rho + 8\rho^2\right)}. \qquad \textbf{(27)}$$

Let $a := r_0^{(2)} + r_0^2\rho/(1-2\rho)$. The condition for T^* to be well defined and positive is

$$4r_0^2\rho - 3r_0^2 + a4(1-\rho)(1-2\rho) > r_0^2.$$

This is true if and only if

$$a(1-2\rho) > r_0^2,$$

which is easily seen to be equivalent to what we stated in (3).

In the deterministic case, $a = r_0^2 + r_0^2\rho/(1-2\rho)$, the condition becomes

$$2\rho < 1 - \frac{r_0^2}{r_0^2 + r_0^2\rho/(1-2\rho)} = 1 - \frac{1}{1+\rho/(1-2\rho)},$$

which can easily be seen to lead to the contradiction $\rho < 0$. ■

3.2. Asymmetric Polling Model with Deterministic Switchover Times

We now consider an a asymmetric polling model (i.e., $\rho_1 > \rho_2$) with deterministic switchover times (i.e., $r_1^{(2)} = r_1^2$ and $r_1^{(2)} = r_1^2$).

PROOF OF THEOREM 4, FIRST PART: Recall that we would like to show that there is no gain from waiting at the station with less traffic (i.e., station 2 in our case). We distinguish two cases: $\rho_1 > 1/2$ and $\rho_1 < 1/2$.

First Case: $\rho_1 > \rho_2$ and $\rho_1 > 1/2$.

Note that trivially $\rho_2 < 1/2$. Recall that the linear relation (25) holds for the minimizers of (24) (which is the same as (2)). Since $\rho_1 > 1/2$, and $\rho_2 < 1/2$, we get $(c_3 - c_4)/(c_5 - 2c_6) < 0$ and $(c_5 - 2c_7)/(c_5 - 2c_6) < 0$. Therefore, due to (25), one of the minimizers T_1^* or T_2^* must be negative. Therefore, the minimizers subject to the restriction $T_1^*, T_2^* \geq 0$ must satisfy either $T_2^* = 0$ or $T_1^* = 0$. However, the second case can be excluded easily: If we set $T_1 = 0$ in (24) and optimize in T_2, we would get

$$T_2^* = -r_0 + \sqrt{r_0^2 + \frac{c_2 - c_4 r_0}{c_7}}.$$

This can be seen to be negative, because

$$c_2 - c_4 r_0 < 0$$

follows from

$$-r_0^2\rho_1 - \frac{\rho_1\rho_2 r_0^2}{1-\rho_0} + \frac{1}{2}\rho_0 r_0^{(2)} = r_0^2\frac{1}{2}(\rho_2 - \rho_1) - \frac{\rho_1\rho_2 r_0^2}{1-\rho_0} < 0,$$

which holds since $\rho_2 < \rho_1$.

Therefore, the case $T_1^* = 0$, $T_2^* > 0$ can be excluded, and we must have $T_1^* \geq 0$ and $T_2^* = 0$ for the minimizers of (24) subject to $T_1^*, T_2^* \geq 0$.

Second Case: $\rho_1 > \rho_2$ and $\rho_1 < 1/2$.

First, let us rewrite the delay formula (2). We exclude the trivial case $r_0 = 0$ and set $S_1 := T_1/r_0$ and $S_2 := T_2/r_0$. Then (2) becomes

$$\bar{D} = c_1 + \frac{r_0}{\rho_0}\left(\frac{\rho_1\rho_2}{1-\rho_0}(1+S_1+S_2) + \left(\frac{1}{2}\rho_0 + S_1\rho_2 + S_2\rho_1 + S_1^2\frac{\rho_2}{2}\frac{1-2\rho_1}{1-\rho_1} + S_2^2\frac{\rho_1}{2}\frac{1-2\rho_2}{1-\rho_2}\right)\Big/(1+S_1+S_2)\right). \quad \textbf{(28)}$$

With the notation S_1 and S_2, the linear relation (25) becomes

$$S_1 = \frac{c_3-c_4}{c_5-2c_6}\frac{1}{r_0} + \frac{c_5-2c_7}{c_5-2c_6}S_2 = (\rho_1-\rho_2)\Big/\left(\rho_2\frac{1-2\rho_1}{1-\rho_1}\right) + \left(\rho_1\frac{1-2\rho_2}{1-\rho_2}\right)\Big/\left(\rho_2\frac{1-2\rho_1}{1-\rho_1}\right)^{-1} S_2.$$

Setting

$$c := (\rho_1-\rho_2)\Big/\left(\rho_2\frac{1-2\rho_1}{1-\rho_1}\right)$$

and

$$b := \left(\rho_1\frac{1-2\rho_2}{1-\rho_2}\right)\Big/\left(\rho_2\frac{1-2\rho_1}{1-\rho_1}\right),$$

this is

$$S_1 = c + bS_2. \quad \textbf{(29)}$$

Consider $\bar{D} = \bar{D}(S_1, S_2)$ (given in (28)) as a function of S_1 and S_2. It suffices to consider the function $\bar{D}(S_1, S_2)$ only at those points that satisfy the linear relation (29)—that is, $\bar{D}(c + bS_2, S_2)$, $S_2 \in [0, \infty)$. We are finished if we can show that the derivative of this function wrt S_2 at 0 is nonnegative, since then the optimum must be attained for negative S_2, and thus negative T_2, which is impossible. Let $g(S_2) := \bar{D}(c + bS_2, S_2)$.

The function g can be written as

$$g(S_2) = c_1 + \frac{r_0}{\rho_0}\left(\frac{\rho_1\rho_2}{1-\rho_0}(1+c+bS_2+S_2) + \left(\frac{1}{2}\rho_0 + (c+bS_2)\rho_2 + S_2\rho_1 + (c+bS_2)^2\frac{\rho_2}{2}\frac{1-2\rho_1}{1-\rho_1} + S_2^2\frac{\rho_1}{2}\frac{1-2\rho_2}{1-\rho_2}\right)\Big/(1+c+bS_2+S_2)\right).$$

Clearly,

$$g'(0) = \frac{r_0}{\rho_0}\left(\frac{\rho_1\rho_2}{1-\rho_0}(1+b) + (1+c)\left(\rho_2 b + \rho_1 + c\rho_2\frac{1-2\rho_1}{1-\rho_1}\,b\right)\right.$$
$$\left.-(1+b)\left(\frac{1}{2}\rho_0 + \rho_2 c + c^2\frac{\rho_2}{2}\frac{1-2\rho_1}{1-\rho_1}\right)\Big/(1+c)^2\right).$$

We would like to show that $g'(0) > 0$, which is true if and only if

$$\frac{\rho_1\rho_2(1+b)(1+c)^2}{1-\rho_0} + \left(\rho_2 b + \rho_1 + c\rho_2\frac{1-2\rho_1}{1-\rho_1}\,b\right)(1+c)$$
$$-\left(\frac{\rho_0}{2} + \rho_2 c + \frac{c^2\rho_2}{2}\frac{1-2\rho_1}{1-\rho_1}\right)(1+b) > 0.$$

After some calculations, it can be seen that this is equivalent to

$$-2\rho_1\rho_2 - \rho_1 + \rho_1^2 + \rho_2 - \rho_2^2 + 2\rho_1\rho_2^2 + 2\rho_1^2\rho_2 < 0,$$

which is easily seen to be always satisfied in the case $\rho_1, \rho_2 < 1/2$. ■

We have seen that in the symmetric polling model with deterministic switchover times there is no gain from waiting at the station with less traffic (station 2). Now, we determine when it is useful to wait at the station with more traffic (station 1) and what is the optimal waiting time T_1^* in this case. It turns out that the condition is (4), the optimal parameter is given by

$$T_1^* = -r_0 + \sqrt{r_0^2 + \frac{c_2 - c_3 r_0}{c_6}}, \tag{30}$$

and the corresponding delay is then obtained by plugging in (30) and $T_2^* = 0$ into (24).

PROOF OF THEOREM 4, SECOND PART: We get the optimal parameter if we set $T_2 = 0$ in (24) and differentiate wrt T_1. Then the minimizer is seen to be given by (30). Condition (4) corresponds to $T_1^* > 0$. In order for (30) to be positive, we must have

$$c_2 - c_3 r_0 > 0,$$

which translates into

$$\frac{\rho_0}{2} > \rho_2 + \frac{\rho_2\rho_1}{1-\rho_0};$$

thus, (4) appears. ■

4. LOWER BOUND

The goal of this section is to give a proof of Theorem 8. For this purpose, let us define the following random variables. We denote by F_i the time in steady state that the server spends at station i waiting in a cycle—that is, being idle because there is no work at that station. Furthermore, let $f_i := \mathbb{E}F_i$, $f_i^{(2)} := \mathbb{E}F_i^2$, and $f_0 := \sum_{i=1}^N f_i$.

For a general strategy basically nothing can be said about the distribution of the F_i even about their means f_i. The idea behind the proof of Theorem 8 is to estimate the mean average delay by an expression in terms of f_i and $f_i^{(2)}$, to estimate by Jensen's inequality,

$$f_i^{(2)} \geq f_i^2, \tag{31}$$

and, thus, to obtain an expression that only depends on the f_i. Then minimizing over $f_i \geq 0$, we obtain the lower bound (6). The details are as follows.

4.1. Decomposition

Let V be the workload as defined above. Analogously to the decomposition principle in (9), one can show that

$$\mathbb{E}V = \mathbb{E}V^{M/G/1} + \mathbb{E}V^{\text{idle}}.$$

We note that in the derivation of (10) and (11), the strategy was not used at all. So, one obtains a lower bound for $\bar{D}$ from a lower bound of $\mathbb{E}V$ and thus from a lower bound for

$$\mathbb{E}V^{\text{idle}} = q\mathbb{E}V^{\text{switching}} + (1-q)\mathbb{E}V^{\text{waiting}}, \tag{32}$$

where $q = \mathbb{P}(\text{server switching}|\text{server idle})$.

Furthermore, we note that the cycle time satisfies

$$(1-\rho_0)\mathbb{E}C = r_0 + f_0.$$

4.2. Workload While Switching

Now, we express $\mathbb{E}V^{\text{switching}}$ in terms of the (unknown) f_i. This is completely analogous to the derivation for our concrete strategy with the arguments following (15) replacing T_i by f_i. The result is

$$\begin{aligned}\mathbb{E}V^{\text{switching}} = {} & \frac{r_0(\rho_0^2 - \sum_{i=1}^N \rho_i^2)}{2\mathbb{P}\,(\text{switching})} + \frac{1}{\mathbb{E}C \cdot \mathbb{P}\,(\text{switching})} \\ & \times \left\{ \rho_0 \frac{r_0^{(2)}}{2} + \sum_{i=1}^N r_i \left[(\rho_0 - \rho_i) f_i + \sum_{j=1}^{i-1} f_j \left(\sum_{l=i+1}^N \rho_l + \sum_{l=1}^{j-1} \rho_l \right) \right.\right. \\ & \left.\left. + \sum_{j=i+1}^N f_j \sum_{l=i+1}^{j-1} \rho_l \right] \right\}. \end{aligned} \tag{33}$$

4.3. Workload While Waiting

Since we do not know the distribution of the waiting time, we cannot say much about the workload generated while the server is waiting. So, we will have to estimate at this point. First, note that

$$\mathbb{E}V^{\text{waiting}} = \frac{1}{\mathbb{P}\,(\text{waiting})} \sum_{i=1}^{N} p_i \mathbb{E}V_i^{\text{waiting}}, \tag{34}$$

where $\mathbb{E}V_i^{\text{waiting}}$ is the expected workload that is present in the system at a point in time when the server is waiting at station i and $p_i = f_i/\mathbb{E}C$.

We cannot calculate the workload present at a point in time when we encounter the server waiting, $\mathbb{E}V_i^{\text{waiting}}$, for an arbitrary strategy, but we *can* give a lower bound; namely we can say that $\mathbb{E}V_i^{\text{waiting}}$ must be at least, on the one hand, the traffic that was accumulated at the other stations during the time that the server has already passed waiting at station i (i.e., the expected backwards recurrence time). Additionally, since the decision of the strategy does not depend on the recent times the server has worked at the other stations or the switchover times, we can also count the terms (16) and (17)—that is the traffic that was accumulated (and is still present) while the server was switching and working at other stations, respectively. This gives

$$\mathbb{E}V_i^{\text{waiting}} \geq \frac{f_i^{(2)}}{2f_i}(\rho_0 - \rho_i) + \sum_{k\neq i} \left((\rho_k \mathbb{E}C + f_k) \sum_{j\in\{i+1,\ldots,k-1\}} \rho_j + r_k \sum_{j\in\{i+1,\ldots,k\}} \rho_j \right).$$

This is the crucial observation in the derivation of the lower bound.

Now, by Jensen's inequality (31), the last term can be yet bounded below by

$$\frac{f_i^2}{2f_i}(\rho_0 - \rho_i) + \sum_{k\neq i} \left((\rho_k \mathbb{E}C + f_k) \sum_{j\in\{i+1,\ldots,k-1\}} \rho_j + r_k \sum_{j\in\{i+1,\ldots,k\}} \rho_j \right). \tag{35}$$

Furthermore, we need that

$$\frac{q}{\mathbb{P}\,(\text{switching})} = \frac{1-q}{\mathbb{P}\,(\text{waiting})} = \frac{1}{\mathbb{P}\,(\text{idle})} = \frac{1}{1-\rho_0}.$$

Then putting (35) back into (34) and this and (33) back into (32) gives a lower bound for $\mathbb{E}V$ (and thus for $\bar{D}$) only in terms of the f_i. Minimizing over the f_i leads to (6).

Note that we cannot count the terms (20) and (21), since, for example, Z_i and F_i are not independent.

5. DIFFERENT STRATEGIES AND OUTLOOK

There is another strategy, which we would like to propose here (we will refer to it as Strategy II), that is likely to be better than the one proposed so far (called Strategy I in

this section) in terms of the mean average delay. However, we are not able to analyze the mean average delay of Strategy II with the present methods.

Strategy II is defined as follows. We consider a polling model as above, the only difference being that the credit T_i now refers to the total time the server spends at station i. More precisely, a server arrives at station i. It then works or waits at station i depending whether messages are present. At time T_i after its arrival at the station, it only finishes all the work that is present at that time (exhaustively). It does not turn idle again at that station in the current cycle (i.e., into "wait-and-see"); whenever the station is empty, it starts switching to the next station. We refer to Figure 3 for an illustration.

It is likely that Strategy II adjusted to its optimal waiting parameters gives a lower average delay than Strategy I adjusted to its optimal waiting parameters $T_1^*, \dots, T_N^*$. Heuristically, Strategy II uses more information about the system, because it also counts the busy periods at the current station. However, we remark that even the determination of the cycle time, as in (7), does not seem to be possible in a straightforward way.

We conjecture that for $N = 2$ and deterministic switchover times, Strategy II provides the lowest mean average delay in the class of all strategies that are not allowed to use information of the queue status at the other station or to look into the future of the system. In this connection, we recall that under the weighting by the traffic intensities ρ_i (rather than the rates λ_i), it is clear that finishing work at the present station (exhaustively) is necessarily better than switching (cf. [10]).

Figure 4 gives a comparison of Strategy I and Strategy II, for which the curve for Strategy II is obtained from simulations. We observed that the optimal credit for Strategy II is *approximated* by $T_1^* + \rho_1 \mathbb{E}C$, where T_1^* is the optimal parameter of Strategy I and $\mathbb{E}C$ is the cycle time of Strategy I for this optimal parameter.

Certainly, one can define different strategies, where, for example, the server additionally has more information on the current queue status at the other stations. This might give an average delay that is even below the lower bound given in Theorem 8. However, note that even if the server is aware of the queue status at all stations, it is not completely clear what is the best decision at each moment in terms of lower average delay: switch or wait-and-see.

Further, one can imagine a situation in which the server may look into the close future of the incoming traffic at the present station, and it might thus decide to abandon

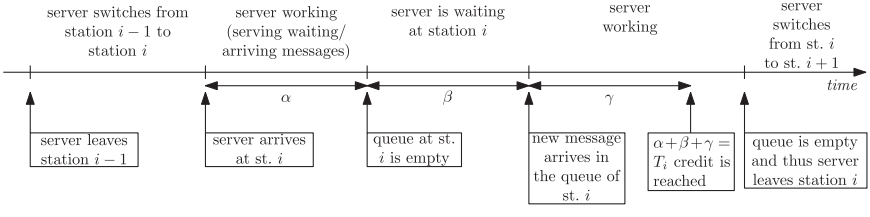

FIGURE 3. Operation of the polling model with Strategy II.

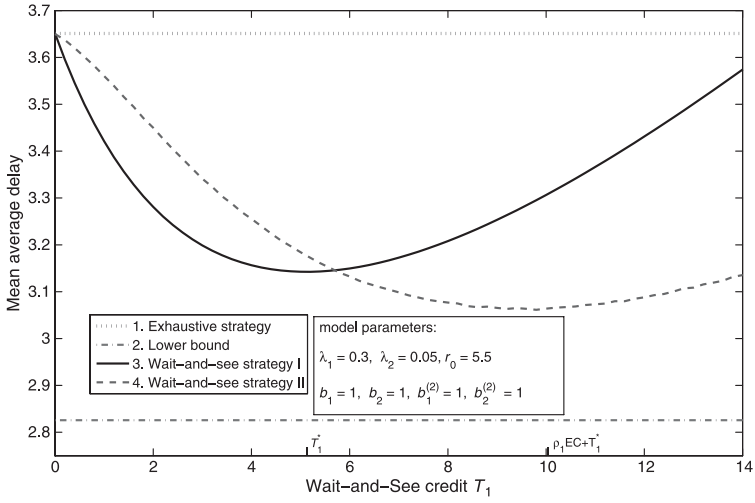

FIGURE 4. Comparison of the optimal credits with Strategy I and Strategy II.

the station before the end of its wait-and-see period, when it is clear that no traffic will arrive during that time.

Acknowledgment

This work was supported by the DFG Research Center "MATHEON – Mathematics for key technologies" in Berlin.

References

1. Boxma, O.J. & Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability* 24(4): 949–964.
2. Boxma, O.J., Levy, H. & Weststrate, J.A. (1991). Efficient visit frequencies for polling tables: minimization of waiting costs. *Queueing Systems* 9: 133–162.
3. Suhov, Yu. M. (ed.), Analytic methods in applied probability. Providence, RI: American Mathematical Society (AMS). Transl., Ser. 2, Am. Math. soc. 207.
4. Boxma, O.J., Schlegel, S. & Yechiali, U. (2002). Two-queue polling models with a patient server. *Annals of Operations Research* 112: 101–121.
5. Cooper, R.B., Niu, S. & Srinivasan, M.M. (1998). When does forced idle time improve performance in polling models? *Management Science* 44: 1079–1086.
6. Cooper, R.B., Niu, S.-C. & Srinivasan, M.M. (1999). Setups in polling models: Does it make sense to set up if no work is waiting? *Journal of Applied Probability* 36(2): 585–592.
7. Kleinrock, L. (1975). *Queueing systems. Volume I*. New York: Wiley.
8. Kramer, G. (2005). *Ethernet passive optical networks*. New York: McGraw-Hill Communications Engineering.
9. Lam, C.F. (2007). *Passive optical networks: Principles and practice*. Amsterdam, Elsevier.
10. Liu, Z., Nain, P. & Towsley, D. (1992). On optimal polling policies. *Queueing Systems*, 11: 59–83.
11. Olsen, T.L. & Van der Mei, R.D. (2003). Polling systems with periodic server routing in heavy traffic: Distribution of the delay. *Journal of Applied Probability* 40(2): 305–326.

12. Peköz, E.A. (1999). More on using forced idle time to improve performance in polling models. *Probability in the Engineering and Informational Sciences* 13(4): 489–496.
13. Takagi, H. (1986). *Analysis of polling models*. Cambridge, MA: MIT Press.
14. Takagi, H. (1988). Queuing analysis of polling models. *ACM Computing Surveys* 20(1): 5–28.
15. Dshalalow, Jewgeni H. (ed.), Frontiers in queueing: models and applications in science and engineering. Boca Raton, FL: CRC Press. Probability and Stochastics Series.
16. Yechiali, U. (2004). On the $M^X/G/1$ queue with a waiting server and vacations. *Sankhyā* 66(1): 159–174.