Probability in the Engineering and Informational Sciences, **32**, 2018, 580–602.
doi:10.1017/S0269964817000419

IMPROVING THE PERFORMANCE OF POLLING MODELS USING FORCED IDLE TIMES

FRANK AURZADA
Fachbereich Mathematik,
Technische Universität Darmstadt,
Darmstadt, Germany
E-mail: aurzada@mathematik.tu-darmstadt.de

SEBASTIAN SCHWINN
Fachbereich Mathematik,
Technische Universität Darmstadt,
Darmstadt, Germany
Graduate School CE,
Technische Universität Darmstadt,
Darmstadt, Germany
E-mail: schwinn@mathematik.tu-darmstadt.de

We consider polling models in the sense of Takagi [19]. In our case, the feature of the server is that it may be forced to wait idly for new messages at an empty queue instead of switching to the next station. We propose four different wait-and-see strategies that govern these waiting periods. We assume Poisson arrivals for new messages and allow general service and switchover time distributions. The results are formulas for the mean average queueing delay and characterizations of the cases where the wait-and-see strategies yield a lower delay compared with the exhaustive strategy.

Keywords: exhaustive service, forced idle time, patient server, polling model, pseudo-conservation law, timer, wait-and-see strategy, waiting time

1. INTRODUCTION

1.1. Model

We investigate a polling model in the sense of [19] consisting of $N \geq 1$ stations, which are served by one server. The stations are labeled by the indices from 1 to N and served in ascending, cyclic order with $N + 1 \triangleq 1$.

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

Switching between stations takes a non-negative random idle time, called the switchover time, where the server does not serve any messages at any station. The random switchover time R_i from station i to the next station (with distribution function F_{R_i}) is assumed to

 0269-9648/17 $25.00

have finite mean r_i and finite second moment $r_i^{(2)}$. We consider both non-deterministic and deterministic switchover times (in the latter case $r_i^{(2)} = r_i^2$ for $i = 1, \ldots, N$). The sum of the mean switchover times is denoted by $r_0 := \sum_{i=1}^{N} r_i$ and the second moment of the sum of independent switchover times by $r_0^{(2)} := \sum_{i=1}^{N} r_i^{(2)} + \sum_{i,j=1, i\neq j}^{N} r_i r_j$.

The message generation process, the lengths of the messages, and the switchover times are assumed to be independent (everything among each other and with respect to the other processes and stations).

The goal of this paper is to obtain explicit formulas for the mean average queueing delay of a message in a polling model with a given wait-and-see strategy in a steady state. The delay is the time a message experiences from the point in time when it arrives in one of the queues until its service starts, that is, excluding the service time. The expected delay of a message generated at station i is denoted by $\mathbb{E}D_i$. The mean average queueing 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 at 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 by the mean message lengths, whereas the delays $\mathbb{E}D_i$ do not include weighting the delay of the individual messages with their lengths. The mean average queueing delay, which we often just abbreviate as *delay*, is called the *intensity weighted mean waiting time* by Takagi [19].

1.2. Wait-and-see strategies

It is characteristic of many service strategies to avoid that the server spends time idly at a station (while there may be work at other stations). In contrast, we focus on wait-and-see strategies where the server may be forced to wait idly for new messages at an empty queue. An advantage is that the server does not switch too often, especially when it is not required or not worthwhile. This can be favorable in order to optimize performance measures such as minimizing the delay, for instance.

First, we describe the wait-and-see behavior of the server in general: The server arrives at a station and starts serving in an exhaustive fashion, that is, serving all waiting messages and newly arriving messages (first come, first served) until the queue is empty. However, once the station is empty or if the server finds an empty station upon its arrival, the server may not immediately switch to the next station; it rather turns idle for some time in order to wait for possibly newly arriving messages ("wait-and-see"). As soon as a new message arrives, the server starts serving immediately and in an exhaustive fashion. Once finished, the server may again turn idle and wait for new messages.

For each of the four strategies considered here, the behavior of the server at station i is governed by a fixed, real parameter $T_i \geq 0$, which has different interpretations (see below). Of course, the server is not allowed to be idle if at its present station messages are waiting to be served. The reason for waiting depends only on the current station in the current cycle, that is, on the evolution of the traffic at the present station since the server arrived there. The server must not use any information about the current queue status at other stations nor about the future of the arrival process at any station. If $T_i = 0$ holds, the service discipline is exhaustive at station i and there is no state of "wait-and-see" at station i. If this is the case for all stations, we call it the *exhaustive strategy*.

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

Now, we specify the four different wait-and-see strategies. Strategy I is extensively analyzed by Aurzada, Beck, and Scheutzow [4] and Strategy IV is examined by Boxma, Schlegel, and Yechiali [6] for $N = 2$ stations and $T_2 = 0$. As far as we know, there are no results in the literature on Strategies II and III.

- Under **Strategy I**, the server has to wait idly the total time T_i for new messages at station i per cycle. Depending on the arrival process, this credit T_i is spent altogether in one single period or in some periods interleaved by different busy periods.
- **Strategy II** is defined as follows: The server has to stay at least the minimum sojourn time T_i at station i per cycle. We can regard it as a timer starting upon arrival of the server at this station. Once the server has spent the minimum sojourn time at the station (possibly consisting of several busy and waiting periods), the server exits the station if the queue is empty. If there are still messages waiting or in service as the timer runs out, the server continues serving in an exhaustive fashion and switches to the next station as soon as the queue is empty.
- **Strategy III** is a modification of the previous one. Here, the server is forced to stay at station i at least the fixed time T_i *after* becoming idle for the first time at this station in this cycle. If there are no messages waiting upon arrival of the server, the timer starts immediately as in the case for Strategy II. Otherwise, the timer starts running just after the first busy period.
- **Strategy IV** is based on a case distinction: If there are messages waiting upon arrival of the server, the server starts serving in an exhaustive fashion and then switches to the next station. On the other hand, if the server finds station i empty upon arrival, a timer is activated and the server remains idle for at most the time T_i, waiting for the first arriving message. If the timer expires before the first arrival occurs, the server switches to the next station. Otherwise, if a new message arrives before the timer expires, the server starts serving immediately and in an exhaustive fashion. After this busy period, the server does not wait any longer at this station in the current cycle and switches to the next station.

We stress that we only deal with strategies where the wait-and-see timers are deterministic. In order to yield a lower minimal delay, we conjecture that deterministic timers do a better job than random timers. Simulations have indicated that such an additional randomness (of the timer) in the polling model has no positive effect on the minimal delay.

1.3. Overview of the contents

The results of this paper are as follows:

- We prove a formula for the mean average queueing delay in a polling model with N stations and Strategy III (Theorem 1).
- We prove a formula for the mean average queueing delay in a polling model with $N = 2$ stations and Strategy II (Theorem 2).
- We extend [6] to timers at *both* stations and prove a formula for the mean average queueing delay for Strategy IV (Theorem 2).
- We characterize the cases for a polling model with $N = 2$ stations where these strategies yield a lower delay compared with the exhaustive strategy (Theorems 3 and 5).

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

The remainder of this paper is structured in the following way: In Section 1.4, we outline related work. Section 2 contains the formulas for the mean average queueing delay (Section 2.1) and the cases where it is worth waiting (Section 2.2). All proofs of the results as well as the determination of essential quantities are collected in Section 3.

1.4. Related work

Aurzada et al. [4] analyze Strategy I and give an explicit formula for the mean average queueing delay in a polling model with N stations. They characterize several cases where Strategy I yields a lower delay compared with the exhaustive strategy. In these cases, the optimal parameters T_i can be computed explicitly. Finally, they give a lower bound for the delay for a class of wait-and-see strategies, which includes Strategies I–IV.

In [6], Boxma et al. focus on a two-queue polling model with a timer as in Strategy IV at station 1, which may be random. They examine different configurations: Either both stations are served exhaustively, or one station is controlled by the 1-limited protocol, whereas the other station is served in an exhaustive fashion. The main results are the probability generating function of the queue lengths, expressions for pseudo-conservation laws, and the Laplace transform of the stationary waiting times.

Besides the main references [4,6], further papers deal with service strategies, which have in common that the server does not necessarily switch to the next station when the current queue is empty. Polling models with deterministic sojourn times and preemptive service are considered in [20] and with exponentially distributed sojourn times in [9]. Similar to Strategy IV, in the setting of [1] the server waits exactly for the first arriving message at an empty station. In [7,16,17], forced idle times are examined where the server is not allowed to resume service immediately as soon as a new message arrives during these idle periods. These three papers are based on an observation in [18], which is as follows: Increasing switchover times can reduce the mean waiting time in polling models.

Furthermore, there are several works that investigate polling models with time-limited service. There, messages are served at a station for a certain period of time or until the queue is empty, whichever occurs first. If there is still work at the station when the timer expires, the server either finishes all the present work, or completes only the service of the currently served message, or stops working immediately at this station and switches to the next station. We refer to [2,8,11,14,15] for random time limits (in particular, exponentially distributed timers). In [10,12], deterministic time limits are studied.

2. RESULTS

In this section, we give formulas for the the mean average queueing delay that allow us to compute the delay for the different wait-and-see strategies. We further characterize the cases where it is favorable (in the sense of a lower delay) to possibly wait at a station instead of switching. From now on, we assume that the stability condition $\rho_0 < 1$ of the polling model holds.

2.1. Main results

Theorems 1 and 2 provide formulas for the mean average queueing delay in terms:

- of the system parameters λ_i, b_i, $b_i^{(2)}$, r_i, $r_i^{(2)}$ for $i = 1, \ldots, N$; and

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

- of the parameter-dependent quantity $\mathbf{S} := (f_i, w_i, \tilde{r}_i)_{i=1,\dots,N}$ of expectations in steady state, which are defined in the next paragraph and which vary depending on the wait-and-see strategy including the parameters T_i. Specifying these expectations for Strategies II–IV in Section 3.2 is the main novelty in this paper.

We define by f_i the expected time per cycle, which the server waits at station i. We use $f_0 := \sum_{i=1}^N f_i$ for the total expected waiting time of the server per cycle (i.e., idle times without switchover times). The expected backward recurrence time (expected spent time) w_i is defined by

$$w_i := \mathbb{E}[\text{time since server arrived at station } i \mid \text{server is idle at station } i],$$

that is, the expectation of the elapsed time since arriving at station i at a random point in time while waiting at station i. Furthermore, we recall the random switchover time R_i from station i to the next station and define the conditional mean switchover time by

$$\tilde{r}_i := \mathbb{E}[R_i \mid \text{server is idle at station } i+1].$$

This means: Given a random point in time while waiting at station $i+1$, the quantity $\tilde{r}_i$ is the expected length of the preceding switchover time.

THEOREM 1: *The mean average queueing delay of a message in a polling model with Strategy III is given by*

$$\begin{aligned}\bar{D} = {} & \frac{\sum_{i=1}^N \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{(r_0+f_0)\left(\rho_0^2 - \sum_{i=1}^N \rho_i^2\right)}{2\rho_0(1-\rho_0)} + \frac{\frac{1}{2}\rho_0 r_0^{(2)} + r_0 \sum_{i=1}^N f_i(\rho_0-\rho_i)}{\rho_0(r_0+f_0)} \\ & + \frac{1}{\rho_0(r_0+f_0)}\left[\sum_{i=1}^N f_i w_i(\rho_0-\rho_i) + \sum_{1\le i<j\le N} f_i f_j(\rho_0-\rho_i-\rho_j)\right] \\ & - \frac{\sum_{i=1}^N f_i\rho_i(\rho_0-\rho_i)}{\rho_0(1-\rho_0)}.\end{aligned} \tag{1}$$

The quantities f_i and w_i for $i = 1, \dots, N$ are specified for exponentially distributed service times in Section 3.2.

We refer to Aurzada et al. [4] for the delay of a message in a polling model with Strategy I. For Strategies II and IV, we restrict the number of stations to $N = 2$ due to the technical effort that would be required otherwise to compute further parameter-dependent quantities, which would arise in the formula for the delay.

THEOREM 2: *The mean average queueing delay of a message in a polling model with $N = 2$ stations and Strategy II as well as Strategy IV is given by*

$$\begin{aligned}\bar{D} = {} & \frac{\sum_{i=1}^2 \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{r_0\rho_1\rho_2}{\rho_0(1-\rho_0)} + \frac{r_0^{(2)}}{2(r_0+f_0)} \\ & + \frac{\rho_2 f_1}{\rho_0(r_0+f_0)}(r_1 + \tilde{r}_2 + w_1) \\ & + \frac{\rho_1 f_2}{\rho_0(r_0+f_0)}(\tilde{r}_1 + r_2 + w_2).\end{aligned} \tag{2}$$

The quantities f_i, w_i and $\tilde{r}_i$ for $i = 1, 2$ are specified in Section 3.2 (in the case of Strategy II only for exponentially distributed service times).

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

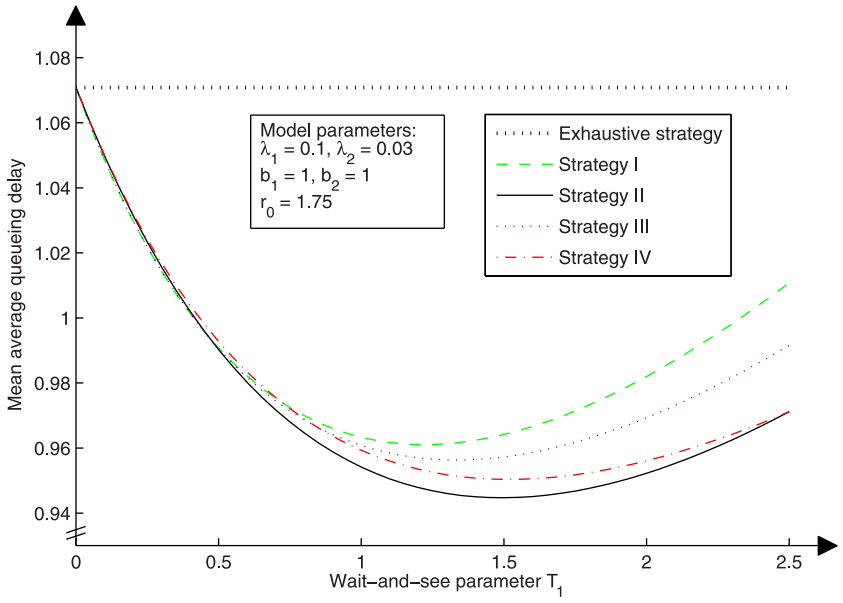

FIGURE **1.** Comparison of the delays for the strategies vs. the wait-and-see parameter T_1.

We stress that formulas (1) and (2) are valid for general distributions of the service times. However, we emphasize that for Strategies II and III we are only able to compute the quantity $\mathbf{S} = (f_i, w_i, \tilde{r}_i)_i$ explicitly for exponentially distributed service times because formula (6) below is only available for the M/M/1 queue in the literature, for instance.

Figure 1 provides a typical relation between the delays for all four wait-and-see strategies. We consider a polling model with $N = 2$ stations where the server is not allowed to wait at station 2. The switchover times are deterministic, symmetrically split among the switchovers and the service times are exponentially distributed. The delay for Strategy I was obtained by the formula from [4], and we used formulas (1) and (2) and the values for $\mathbf{S} = (f_i, w_i, \tilde{r}_i)_i$ from Section 3.2 to compute the data for Strategies II–IV.

The ranking of the wait-and-see strategies with respect to the minimal delay observed in Figure 1 can be explained naturally: In the best case, the server exits the current station as soon as there is enough work waiting at the other station. Since the server does not have any information about the queue status at the other station, the sojourn time at the current station is the crucial quantity in order to estimate the workload generated at the other station. Hence, there is an optimal sojourn time for each station, and the minimal delay is attained if the expected sojourn time agrees with the optimal sojourn time as good as possible, that is, with small variance.

Therefore, we conjecture that Strategy III always yields a lower minimal delay compared with Strategy I and that Strategy II is the best of the investigated wait-and-see strategies.

How much can be gained by wait-and-see strategies varies depending on the system parameters. It is even possible to give examples where the relation between the delay in a polling model with forced idle times and the delay for the exhaustive strategy is arbitrarily small.

2.2. Is it worth waiting?

In addition to the formulas for the delay, Theorems 1 and 2 allow us to put the following question: Given the system parameters, how does one have to adjust the parameters $T_i \geq 0$

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

such that the delay is minimized. In this context, we regard the delay $\bar{D}$ for $N = 2$ as a function of the T_i and thus write $\bar{D}(T_1, T_2)$. In general, we cannot compute a minimizer of this problem

$$\min_{T_1 \geq 0,\, T_2 \geq 0} \bar{D}(T_1, T_2)$$

for Strategies II–IV analytically. Nevertheless, we do obtain necessary and sufficient conditions for these wait-and-see strategies in a polling model with exponentially distributed service times such that it is favourable to wait in comparison with the exhaustive strategy, that is, there exists $(T_1, T_2) \neq (0, 0)$ such that $\bar{D}(T_1, T_2) < \bar{D}(0, 0)$.

Note that we only consider the two cases with the additional restriction $T_2 = 0$ and $T_1 = T_2$, respectively. As a summary, one can say that the benefit of waiting arises from the asymmetry of the system or from non-deterministic switchover times.

THEOREM 3: *Let $N = 2$ and $T_2 = 0$. It is worth waiting at station 1, that is, there exists $T_1 > 0$ such that $\bar{D}(T_1, 0) < \bar{D}(0, 0)$, in a polling model with*

- *Strategy III if and only if*

$$\frac{r_0^{(2)}}{2r_0^2} - \frac{\rho_2(1-\rho_2)}{\rho_0(1-\rho_0)} > 0.$$

- *Strategy II as well as Strategy IV if and only if*

$$\frac{r_0^{(2)}}{2r_0\left(r_1 + \tilde{r}_2^{\mathrm{IV}}\right)} - \frac{\rho_2}{\rho_0} > 0, \tag{3}$$

where the quantity $\tilde{r}_2^{\mathrm{IV}}$ is defined just above Theorem 1 and can be computed as in (18) below. In the case of a deterministic switchover time R_2, inequality (3) simplifies to $\rho_1 > \rho_2$.

We mentioned above that the parameter-dependent quantities f_i, w_i and $\tilde{r}_i$ can vary depending on the wait-and-see strategy including the parameters T_i. The dependence on the strategy is indicated by a superscript if necessary, for example, $\tilde{r}_2^{\mathrm{IV}}$. However, note that condition (3), which contains $\tilde{r}_2^{\mathrm{IV}}$, *must not* depend on the T_i. Indeed, the quantity $\tilde{r}_2^{\mathrm{IV}}$ does not depend on the T_i (see (18) below).

Remark 4: First, we introduce the coefficient of variation c_X of a random variable X, which is defined as the ratio of the standard deviation and the mean. Furthermore, we denote by R_0 the sum of the switchover times. We can observe from Theorem 3 that for c_{R_0} sufficiently large, it is even worth waiting at station 1 in spite of a lower traffic load $\rho_1 < \rho_2$. As a consequence of this, we can make a conjecture for a polling model with $N = 2$ stations and without any restriction on the T_i: In the case of c_{R_0} sufficiently large, it is favorable to have *positive* parameters T_i at *both* stations instead of just allowing "wait-and-see" at the station with higher traffic load.

Similar to above, we get necessary and sufficient conditions for a symmetric polling model with $\rho_1 = \rho_2$ and the restriction $T_1 = T_2$ such that the delay is lower than for the exhaustive strategy. The arrival rates, message length and switchover time distributions are also assumed to be the same for both stations for Strategies II and IV, but we can omit this requirement for Strategy III.

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

THEOREM 5: *Let $N = 2$. It is worth waiting with the restriction $T_1 = T_2$, that is, there exists $T_1 > 0$ such that $\bar{D}(T_1, T_1) < \bar{D}(0, 0)$, in a symmetric polling model with*

- *Strategy III if and only if*

$$\frac{r_0^{(2)}}{r_0^2} - \frac{1-\rho_1}{1-\rho_0} > 0$$

  (that can only be satisfied for non-deterministic switchover times),
- *Strategy II as well as Strategy IV if and only if the switchover times are non-deterministic.*

We give a direct consequence of the two preceding theorems:

COROLLARY 6: *There are parameter settings of a polling model with $N = 2$ stations where Strategies II and IV yield a lower delay than Strategies I and III, that is, it is only worth waiting with Strategies II and IV.*

3. PROOFS

3.1. Proofs of the main results

We show how to derive Theorems 1 and 2 which are based on a decomposition principle from [5] and on the technique of the proofs of Theorems 1 and 8 from [4]. We mention that the proofs of Theorems 1 and 2 are quite standard and that the key novelty is the computation of the parameter-dependent quantities in Section 3.2.

PROOF OF THEOREM 1: First, we recall some important identities: The cycle time is the time that the server takes from its arrival at station 1 to the next arrival at this station. The mean cycle time in steady state is denoted by $\mathbb{E}C$ and is given by

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

Indeed, this can be argued by looking at the expected time which the server is idle per cycle. This expectation equals the sum of all mean switchover and waiting times, that is, $r_0 + f_0$. On the other hand, we can represent this expected idle time using the total load offered to the system which leads to $(1-\rho_0)\mathbb{E}C$.

Next, we refer to a workload decomposition principle in [5,6] which also holds for our system. We omit a proof since it is analogous to the proofs there. As a consequence of this decomposition principle, we obtain

$$\mathbb{E}V = \mathbb{E}V^{\mathrm{M/G/1}} + q\mathbb{E}V^{\mathrm{switching}} + (1-q)\mathbb{E}V^{\mathrm{waiting}}, \tag{4}$$

where $q \coloneqq \mathbb{P}(\text{server is switching} \mid \text{server is idle})$ and V is the workload at a random point in time in steady state. The workload consists of the sum of all message lengths that are present in the system including the remaining service time of the currently served message. The quantities $V^{\mathrm{M/G/1}}$ and $V^{\mathrm{switching}}$ (V^{waiting}) refer to the workload in the same polling model without switchover and waiting times, and to the workload given that the server is

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

switching (waiting) at a random point in time, respectively. Furthermore, we can determine the expected workload differently by

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

where the first sum accounts for the messages that are not yet in service and the second sum for the currently served message. The quotient is the expected residual service time of a currently served message at station i. Using Little's law, this equation can be rearranged into

$$\mathbb{E}V = \rho_0 \bar{D} + \sum_{i=1}^{N} \rho_i \frac{b_i^{(2)}}{2b_i}. \tag{5}$$

Therefore, we can combine (4) and (5) in order to obtain a representation of the delay $\bar{D}$. The quantity $\mathbb{E}V^{\mathrm{M/G/1}}$ is given in the literature, for example, in [3, p. 206].

From now on, we focus on the expected workload present while switching $\mathbb{E}V^{\text{switching}}$ and waiting $\mathbb{E}V^{\text{waiting}}$. The former does not directly depend on the given wait-and-see strategy so that we can proceed in the same way as in [4]. On the other hand, the particular wait-and-see strategy influences the expected workload present while waiting. It remains to give the general formula for $\mathbb{E}V_i^{\text{waiting}}$, the expected workload that is present in the system at a random point in time when the server is waiting at station i. Following the computation in [4] [see Eq. (23) there], we obtain

$$\begin{aligned}
\mathbb{E}V_i^{\text{waiting}} = {} & \sum_{j<i} r_j \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j} \rho_l \right) + \sum_{j>i} r_j \sum_{l=i+1}^{j} \rho_l \\
& + \sum_{j<i} \rho_j \mathbb{E}C \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j-1} \rho_l \right) + \sum_{j>i} \rho_j \mathbb{E}C \sum_{l=i+1}^{j-1} \rho_l \\
& + \sum_{j<i} f_j \left(\sum_{l=i+1}^{N} \rho_l + \sum_{l=1}^{j-1} \rho_l \right) + \sum_{j>i} f_j \sum_{l=i+1}^{j-1} \rho_l \\
& + (\rho_0 - \rho_i) w_i,
\end{aligned}$$

where w_i denotes the expectation of the elapsed time since arriving at station i at a random point in time while waiting at station i. Combining all the relevant equations, we get the formula in Theorem 1 for the delay. ■

Proof of Theorem 2: For $N = 2$, the only part that differs from the proof of Theorem 1 is the computation of $\mathbb{E}V_i^{\text{waiting}}$ for $i = 1, 2$. Now, we focus on this quantity for $i = 1$ and thus have to keep the condition in mind that the server is currently waiting at station 1. Let us assume that the server is at such a point in time. Then, the present workload has not been generated at station 1 (otherwise the server would not be waiting at this station). Therefore, the workload which is currently present can only consist of messages that have been generated at station 2 since exiting that station. The expectation of the elapsed time since exiting station 2 is the sum of the conditional mean switchover time $\tilde{r}_2$ and the expected backward recurrence time w_1 whose definitions can be found just above Theorem 1. We get

$$\mathbb{E}V_1^{\text{waiting}} = \rho_2(\tilde{r}_2 + w_1)$$

and for $\mathbb{E}V_2^{\text{waiting}}$, we just have to exchange the roles of 1 and 2. ■

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

Remark 7: Actually, the conditional mean switchover time $\tilde{r}_i$ only differs from r_i for a non-deterministic switchover time for Strategies II and IV. In the case of deterministic switchover times for Strategies II and IV, or in the case of a polling model with Strategies I and III (waiting occurs if $T_i > 0$, independently of the switchover times), we have the equality $\tilde{r}_i = r_i$.

Remark 8: We briefly refer to Theorem 8 in [4] which provides a lower bound for the delay for a class of wait-and-see strategies (including Strategies I–IV). The bound given there is correct for a polling model with $N = 2$ stations and deterministic switchover times. In the case of $N = 2$ and non-deterministic switchover times, we can replace r_k [in (35) there] by the mean of the switchover time R_k from station k to the next station given that there is no message arriving at station $k + 1$ while switching. For instance, this quantity equals $\mathbb{E}[R_2 \mid B_0]$ in (18) below for $k = 2$. In addition for a polling model with $N > 2$ stations, one has to give lower bounds for further quantities, which arise in the proof there.

3.2. Determination of $\mathbf{S} = (f_i, w_i, \tilde{r}_i)_i$

The general formulas for the delay in Theorems 1 and 2 require the specification of $\mathbf{S} = (f_i, w_i, \tilde{r}_i)_i$ according to the wait-and-see strategy. We recall that the definitions of these quantities can be found just above Theorem 1. The real novelty of this work is the determination of these quantities in this section. Note that we restrict the service times of the messages to exponential distributions with parameter $\mu_i := 1/b_i$ at station i for $i = 1, \ldots, N$ for Strategies II and III. After the following preparations, we discuss the different wait-and-see strategies separately where some terms from [6] can also be spotted here for Strategy IV. For the sake of simplicity, we deal with Strategy III before Strategy II.

3.2.1. Preparations. It is helpful to introduce c_i the expected time per cycle in steady state, which the server spends at station i. This expression is directly related to the mean cycle time $\mathbb{E}C$ and to the expected waiting time f_i at station i by the equation

$$c_i = \rho_i \mathbb{E}C + f_i.$$

Moreover, we define $c_0 := \sum_{i=1}^{N} c_i$ and obtain $\mathbb{E}C = c_0 + r_0$.

We require a time-dependent state probability [denoted by $P_{j,k}(x)$] to analyze the delay for Strategies II and III, and we require the distribution of the length of a busy period to analyze the delay for Strategies II and IV.

The probability $P_{j,k}(x)$. According to [13, pp. 53–78], we denote by $P_{j,k}(x)$ the probability that the queue length (including the possibly current service) of an $M/M/1$ queue with arrival rate λ and service rate μ is k at time x given that the queue length is j at time zero. We introduce the abbreviation $a := 2\mu\sqrt{\rho}$, where the traffic load ρ equals λ/μ, and the modified Bessel functions $I_k(x)$ of the first kind of order k, which can be defined by

$$I_k(x) := \sum_{m=0}^{\infty} \frac{\left(\frac{x}{2}\right)^{k+2m}}{(k+m)!m!} \quad \text{for } k \in \mathbb{N}_0$$

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

and $I_{-k}(x) := I_k(x)$ for $k \in \mathbb{N}$. Finally, we have

$$P_{j,k}(x) = e^{-(\lambda+\mu)x}\Bigg[\rho^{\frac{k-j}{2}} I_{k-j}(ax) + \rho^{\frac{k-j-1}{2}} I_{k+j+1}(ax) + (1-\rho)\rho^k \sum_{l=k+j+2}^{\infty} \rho^{-\frac{l}{2}} I_l(ax)\Bigg] \tag{6}$$

due to [13, p. 77]. We emphasize that we have to set $\lambda := \lambda_i$ and $\mu := \mu_i$ if we focus on station i. Hence, the probability $P_{j,k}(x)$ can differ depending on i, but we omit such an additional index because it arises out of the context.

The density g_i of a busy period. The density of the length of a busy period at station i is denoted by g_i and the n-fold convolution of g_i with itself by $g_i^{(*n)}$. We get

$$g_i(x) = \sum_{n=1}^{\infty} e^{-\lambda_i x} \frac{(\lambda_i x)^{n-1}}{n!} b_i^{(*n)}(x) \quad \text{for } x \geq 0$$

from [13, p. 226]. Note that with abuse of notation $g_i^{(*0)}$ represents the Dirac delta function according to the property that the length of 0 busy periods is zero. The density $b_i^{(*n)}$ is the n-fold convolution of the service time with itself. For exponentially distributed service times, we obtain the density

$$b_i^{(*n)}(x) = \frac{\mu_i^n x^{n-1}}{(n-1)!} e^{-\mu_i x} \quad \text{for } x \geq 0,$$

of the Erlang(n, μ_i) distribution, which can also be identified as a gamma distribution. In this particular case, a further representation of g_i using the modified Bessel function of the first kind of order one is given in [13, p. 215].

3.2.2. Strategy III. We denote by $q_i(T_i)$ the expected number of messages (including the possibly currently served message) present at station i after time T_i given that there is no message present at time zero. With the probability $P_{0,k}(T_i)$ from (6), we get

$$q_i(T_i) = \sum_{k=0}^{\infty} k P_{0,k}(T_i).$$

Since we only require the expected number of messages at time T_i, we define the short version $q_i := q_i(T_i)$.

The expected sojourn time c_i. For each station we get the equation

$$c_i = \lambda_i(r_0 + c_0 - c_i)\frac{b_i}{1-\rho_i} + T_i + q_i \frac{b_i}{1-\rho_i}, \tag{7}$$

which can be seen as follows: First of all, the time which the server spends at station i depends on the elapsed time since exiting this station in the preceding cycle up to the current arrival at this station. This expected intervisit time of the server at station i is

$$\mathbb{E}C - c_i = r_0 + c_0 - c_i$$

and the quotient $b_i/(1-\rho_i)$ is the expected length of a busy period (which is caused by one arriving message). In order to obtain this latter quantity, we refer to the short calculation

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

using Laplace transforms in [13, pp. 211–213]. Together with the arrival rate λ_i, we can compute the expected length of the first busy period (generated by the waiting messages) at station i and get

$$\lambda_i(r_0 + c_0 - c_i)\frac{b_i}{1-\rho_i}. \tag{8}$$

After the first busy period, the server has to spend the time T_i at this station (which can consist of several busy and waiting periods). Then, the server exits the station if the queue is empty at time T_i. Alternatively, if there are messages present at time T_i, the server continues serving messages until the queue is empty. This additional time depends on the expected number q_i of present messages and equals $q_i b_i/(1-\rho_i)$ in expectation.

Using (7), we can set up a linear system of equations with variables c_i. For instance, in the case of two stations, we obtain

$$\begin{aligned} c_1 &= \frac{r_0\rho_1 + (1-\rho_1)(1-\rho_2)T_1 + \rho_1(1-\rho_2)T_2 + (1-\rho_2)q_1b_1 + \rho_1q_2b_2}{1-\rho_0}, \\ c_2 &= \frac{r_0\rho_2 + (1-\rho_1)(1-\rho_2)T_2 + \rho_2(1-\rho_1)T_1 + (1-\rho_1)q_2b_2 + \rho_2q_1b_1}{1-\rho_0}. \end{aligned} \tag{9}$$

The expected backward recurrence time w_i. The expectation w_i is the sum of two terms: On the one hand, there is the expected length of the first busy period at station i [see term (8)]. The second summand is the expectation of the elapsed time since becoming idle at station i for the first time at a random point in time while waiting at this station. Therefore, we get

$$w_i = \lambda_i(r_0 + c_0 - c_i)\frac{b_i}{1-\rho_i} + \frac{\int_0^{T_i} xP_{0,0}(x)\,dx}{\int_0^{T_i} P_{0,0}(x)\,dx}, \tag{10}$$

where a random point in time while waiting has the density

$$\frac{P_{0,0}(x)}{\int_0^{T_i} P_{0,0}(y)\,dy} \quad \text{for } x \in [0, T_i].$$

3.2.3. Strategy II. We focus on the steady-state probabilities $\pi_n^{(i)}$ for all $n \in \mathbb{N}_0$ that the server finds n messages waiting upon arrival at station i. We consider deterministic switchover times in this paragraph first. The following system of equations describes the relation of consecutive visits at the stations. The probability of finding n messages upon arrival at station 1 depends on the intervisit time of the server, that is, the time since exiting this station in the preceding cycle. The intervisit time can be divided into the sum of the switchover times and the time which the server spends at station 2 between two consecutive visits at station 1. This latter time can be split in two parts: First, the server stays the minimum sojourn time T_2. The second part consists of the time which the server takes to serve the possibly remaining messages. This part depends on the number of messages present at time T_2. Given that the server finds k messages upon arrival at station 2, there are l messages present with probability $P_{k,l}(T_2)$ after spending the minimum sojourn time. Then, the length of the second part has the density $g_2^{(*l)}$ which denotes the density of the sum of l independent busy periods at station 2. We recall that the arrival process at station 1 is a Poisson process with arrival rate λ_1. The probability of finding n messages at station 1

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

is given by a Poisson distribution with parameter $\lambda_1 t$ if the intervisit time of the server equals t. Therefore, we can conclude the equation

$$\pi_n^{(1)} = \sum_{k=0}^{\infty} \pi_k^{(2)} \sum_{l=0}^{\infty} P_{k,l}(T_2) \int_0^{\infty} e^{-\lambda_1(r_0+T_2+x)} \frac{(\lambda_1(r_0+T_2+x))^n}{n!} g_2^{(*l)}(x)\,dx$$

for deterministic switchover times. Thereby, we get the coefficients for an infinite linear system of equations $\pi^{(1)} = A\pi^{(2)}$. In the same manner as above, there is a system of equations $\pi^{(2)} = B\pi^{(1)}$.

If the switchover times are non-deterministic, we cannot proceed in such a straightforward way. Instead, we focus on the queue length distribution at server departure instants. Note that the queue at departure instants is always empty at the current station. We denote by $\nu_n^{(i)}$ the steady-state probabilities that there are n messages waiting at the other station upon exit from station i. Now, we give an explanation for the equation

$$\begin{aligned}\nu_n^{(1)} = \sum_{k=0}^{\infty} \nu_k^{(2)} \sum_{m=0}^{\infty} \sum_{j=0}^{n} \Bigg[\int_0^{\infty} e^{-(\lambda_1+\lambda_2)x} \frac{(\lambda_1 x)^m}{m!} \frac{(\lambda_2 x)^j}{j!}\, dF_{R_2}(x) \sum_{l=0}^{\infty} P_{k+m,l}(T_1) \\ \times \int_0^{\infty} e^{-\lambda_2(T_1+x)} \frac{(\lambda_2(T_1+x))^{n-j}}{(n-j)!} g_1^{(*l)}(x)\,dx \Bigg],\end{aligned} \tag{11}$$

which consists of similar terms as above. Given that there are k messages waiting at station 1 upon exit from station 2, we have m message arrivals at station 1 and j message arrivals at station 2 while switching to station 1. Therefore, there are $k+m$ messages waiting at station 1 upon arrival at this station. In order to obtain a queue length of n messages at station 2 upon exit from station 1, a total of $n-j$ messages have to arrive at station 2 during this stay. Then, equation (11) follows by considering all possible variations of indices.

From (11) and the corresponding observation, we get two linear systems of equations $\nu^{(1)} = \tilde{A}\nu^{(2)}$ and $\nu^{(2)} = \tilde{B}\nu^{(1)}$ where the coefficients of $\tilde{A}$ are given in (11). Finally, we are able to determine $\pi_n^{(i)}$ by

$$\pi_n^{(1)} = \sum_{k=0}^{n} \nu_k^{(2)} \int_0^{\infty} e^{-\lambda_1 x} \frac{(\lambda_1 x)^{n-k}}{(n-k)!}\, dF_{R_2}(x). \tag{12}$$

For $\pi^{(2)}$, the roles of 1 and 2 have to be exchanged.

The expected sojourn time c_i. Using the solutions $\pi^{(i)}$, we obtain the expected sojourn time

$$c_i = T_i + \sum_{k=0}^{\infty} \pi_k^{(i)} \sum_{l=0}^{\infty} l P_{k,l}(T_i) \frac{b_i}{1-\rho_i},$$

which the server spends at station i per cycle. Here, the series

$$\sum_{l=0}^{\infty} l P_{k,l}(T_i)$$

is the expectation of the number of messages present at station i after spending the minimum sojourn time T_i given that there are k messages present upon arrival of the server. The quotient $b_i/(1-\rho_i)$ is the expected length of a busy period.

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

The expected backward recurrence time w_i. In order to determine w_i, we recall the condition that a point in time while the server is waiting is randomly chosen. We distinguish how many messages are waiting upon arrival of the server at the station. Therefore, we obtain

$$w_i = \sum_{k=0}^{\infty} p_k^{(i)} \frac{\int_0^{T_i} x P_{k,0}(x)\,dx}{\int_0^{T_i} P_{k,0}(x)\,dx},$$

where $p_k^{(i)}$ denotes the probability of choosing a waiting period during a stay with k messages waiting upon arrival of the server. Similar to Strategy III above, the quotient is the expectation of the elapsed time since arriving at station i at a random point in time while waiting at station i given that there are k messages waiting upon arrival of the server.

It remains to determine the coefficients $p_k^{(i)}$. The basic observation is that $p_k^{(i)}$ is proportional to the product of the probability $\pi_k^{(i)}$ that the server finds k messages waiting upon arrival at station i and the expected length of the total waiting time during the stay at such a station, that is, $\int_0^{T_i} P_{k,0}(x)\,dx$. Hence, the probability $p_k^{(i)}$ is given by

$$p_k^{(i)} = \frac{\pi_k^{(i)} \int_0^{T_i} P_{k,0}(x)\,dx}{\sum_{l=0}^{\infty} \pi_l^{(i)} \int_0^{T_i} P_{l,0}(x)\,dx}. \tag{13}$$

The conditional mean switchover time $\tilde{r}_i$. If the switchover time from station i to the next station is deterministic, we get $\tilde{r}_i = r_i$ (cf. Remark 7). Otherwise, the conditional mean switchover time $\tilde{r}_i$ from station i to the next station, given a random point in time while waiting at station $i+1$, can be determined as follows. We restrict the computation to $i = 2$ for the sake of clarity. First, we introduce the events

$A_l \coloneqq \{$there are l messages waiting at station 1 upon exit from station 2$\}$,

$B_j \coloneqq \{$there are j messages arriving at station 1 while switching from station 2 to 1$\}$,

$C_k \coloneqq \{$there are k messages waiting at station 1 upon arrival$\}$

for all $j, k, l \in \mathbb{N}_0$. We distinguish how many messages are waiting upon arrival of the server at station 1 just like above. We get

$$\tilde{r}_2 = \sum_{k=0}^{\infty} p_k^{(1)}\, \mathbb{E}[R_2 \mid C_k], \tag{14}$$

where $p_k^{(i)}$ is given by (13). Now, we are left with the specification of the quantity $\mathbb{E}[R_2 \mid C_k]$. We make use of

$$C_k = \bigcup_{j=0}^{k} A_{k-j} \cap B_j$$

and obtain

$$\mathbb{E}[R_2 \mid C_k] = \sum_{j=0}^{k} \frac{\mathbb{P}(A_{k-j} \cap B_j)}{\mathbb{P}(C_k)}\, \mathbb{E}[R_2 \mid A_{k-j} \cap B_j].$$

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

Due to the independence of the events A_{k-j} and B_j, and the fact that A_{k-j} does not influence the switchover time R_2, we get

$$\mathbb{E}[R_2 \mid C_k] = \sum_{j=0}^{k} \frac{\mathbb{P}(A_{k-j})\mathbb{P}(B_j)}{\mathbb{P}(C_k)} \mathbb{E}[R_2 \mid B_j]. \tag{15}$$

It remains to determine these quantities. We can represent the event B_j as

$$B_j = \left\{ \sum_{l=1}^{j} e_l \le R < \sum_{l=1}^{j+1} e_l \right\}, \tag{16}$$

where $(e_l)_l$ is a sequence of independent and exponentially distributed random variables with parameter 1 which are independent of $R \coloneqq \lambda_1 R_2$ as well. We get

$$\lambda_1 \mathbb{E}\left[R_2 \mid B_j\right] = \frac{\mathbb{E}\left[R \mathbb{1}_{B_j}\right]}{\mathbb{E}\left[\mathbb{1}_{B_j}\right]} = \frac{\mathbb{E}_R\left[R \mathbb{E}_{(e_l)_l}\left[\mathbb{1}_{B_j}\right]\right]}{\mathbb{E}_R\left[\mathbb{E}_{(e_l)_l}\left[\mathbb{1}_{B_j}\right]\right]}.$$

We use the property that the sum of independent and identically exponentially distributed random variables is Erlang distributed and thus compute

$$\mathbb{E}_{(e_l)_l}\left[\mathbb{1}_{B_j}\right] = \frac{R^j}{j!} e^{-R}.$$

Therefore, we obtain

$$\mathbb{E}[R_2 \mid B_j] = \frac{\mathbb{E}_R\left[R^{j+1} e^{-R}\right]}{\lambda_1 \mathbb{E}_R\left[R^j e^{-R}\right]} = \frac{\int_0^\infty x e^{-\lambda_1 x} \frac{(\lambda_1 x)^j}{j!}\, dF_{R_2}(x)}{\int_0^\infty e^{-\lambda_1 x} \frac{(\lambda_1 x)^j}{j!}\, dF_{R_2}(x)} \tag{17}$$

and

$$\mathbb{P}(B_j) = \mathbb{E}\left[\mathbb{1}_{B_j}\right] = \int_0^\infty e^{-\lambda_1 x} \frac{(\lambda_1 x)^j}{j!}\, dF_{R_2}(x).$$

Finally, we have

$$\mathbb{P}(C_k) = \sum_{j=0}^{k} \mathbb{P}(A_{k-j})\mathbb{P}(B_j)$$

due to the independence and $\mathbb{P}(A_{k-j}) = \nu_{k-j}^{(2)}$.

3.2.4. Strategy IV. As above, $\pi_n^{(i)}$ is the steady-state probability that the server finds n messages waiting upon arrival at station i. The method we use to give the characterizing system coincides with the method for Strategy II. The probability $\pi_n^{(1)}$ depends on the intervisit time of the server, which consists of the switchover times and the time that the server spends at station 2 between two consecutive visits at station 1.

We have to distinguish whether there is no message or at least one message waiting at station 2 because it influences the activation of the timer. In the first case, either a new message arrives before the timer expires and a busy period starts, or there is no message

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

arrival and the server waits the whole time T_2. For deterministic switchover times, we obtain

$$\pi_n^{(1)} = \pi_0^{(2)} \left[\int_0^{T_2} \int_0^{\infty} e^{-\lambda_1(r_0+x+y)} \frac{(\lambda_1(r_0+x+y))^n}{n!} g_2(x)\,dx\,\lambda_2 e^{-\lambda_2 y}\,dy \right.$$
$$\left. + e^{-\lambda_1(r_0+T_2)} \frac{(\lambda_1(r_0+T_2))^n}{n!} e^{-\lambda_2 T_2} \right]$$
$$+ \sum_{k=1}^{\infty} \pi_k^{(2)} \int_0^{\infty} e^{-\lambda_1(r_0+x)} \frac{(\lambda_1(r_0+x))^n}{n!} g_2^{(*k)}(x)\,dx.$$

Once again, we get systems of equations $\pi^{(1)} = A\pi^{(2)}$ and $\pi^{(2)} = B\pi^{(1)}$. Note that we are only interested in $\pi_0^{(i)}$ in the end.

In the case of non-deterministic switchover times, we focus on the steady-state probabilities $\nu_n^{(i)}$ that there are n messages waiting at the other station upon exit from station i. We obtain

$$\nu_n^{(1)} = \nu_0^{(2)} \sum_{j=0}^{n} \left[\int_0^{\infty} e^{-(\lambda_1+\lambda_2)x} \frac{(\lambda_1 x)^0}{0!} \frac{(\lambda_2 x)^j}{j!}\,dF_{R_2}(x) \right.$$
$$\times \left(\int_0^{T_1} \int_0^{\infty} e^{-\lambda_2(x+y)} \frac{(\lambda_2(x+y))^{n-j}}{(n-j)!} g_1(x)\,dx\,\lambda_1 e^{-\lambda_1 y}\,dy \right.$$
$$\left. \left. + e^{-\lambda_2 T_1} \frac{(\lambda_2 T_1)^{n-j}}{(n-j)!} e^{-\lambda_1 T_1} \right) \right]$$
$$+ \sum_{k=0}^{\infty} \nu_k^{(2)} \sum_{\substack{m=0 \\ m+k\neq 0}}^{\infty} \sum_{j=0}^{n} \left[\int_0^{\infty} e^{-(\lambda_1+\lambda_2)x} \frac{(\lambda_1 x)^m}{m!} \frac{(\lambda_2 x)^j}{j!}\,dF_{R_2}(x) \right.$$
$$\left. \times \int_0^{\infty} e^{-\lambda_2 x} \frac{(\lambda_2 x)^{n-j}}{(n-j)!} g_1^{(*(k+m))}(x)\,dx \right]$$

and get two systems of equations $\nu^{(1)} = \tilde{A}\nu^{(2)}$ and $\nu^{(2)} = \tilde{B}\nu^{(1)}$. Finally, we can compute $\pi_n^{(i)}$ as mentioned in (12) for Strategy II.

The expected waiting time f_i. Let E_i be an exponentially distributed random variable with intensity λ_i, which represents the interarrival time of messages at station i. We denote by $\min(E_i, T_i)$ the random length of a waiting period at station i. The timer at station i is activated if and only if the server finds this station empty upon arrival. Therefore, we can conclude

$$f_i = \pi_0^{(i)} \mathbb{E}\left[\min(E_i, T_i)\right] = \frac{\pi_0^{(i)}}{\lambda_i} \left(1 - e^{-\lambda_i T_i}\right)$$

for the expected waiting time at station i per cycle in steady state.

The expected backward recurrence time w_i. The quantity w_i equals the expected residual time of a waiting period and is given by

$$w_i = \frac{\mathbb{E}\left[\min(E_i, T_i)^2\right]}{2\mathbb{E}\left[\min(E_i, T_i)\right]} = \frac{1}{\lambda_i} - \frac{T_i}{e^{\lambda_i T_i} - 1}.$$

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

The conditional mean switchover time $\tilde{r}_i$. If the switchover time is deterministic, we just have $\tilde{r}_i = r_i$ (cf. Remark 7). Now, we focus on a non-deterministic switchover time: Similar but easier than for Strategy II, the quantity $\tilde{r}_2$ is just the mean switchover time given that there is no arrival at station 1 while switching to this station. We get

$$\tilde{r}_2 = \mathbb{E}[R_2 \mid B_0] = \frac{\int_0^\infty x e^{-\lambda_1 x}\, dF_{R_2}(x)}{\int_0^\infty e^{-\lambda_1 x}\, dF_{R_2}(x)} \tag{18}$$

and we can represent $\tilde{r}_1$ in an analogous manner.

3.3. Proofs of the "worth-waiting" results

3.3.1. Preparations. First, we state two facts which we use later to prove that it is worth waiting with Strategy II if it is worth waiting with Strategy IV. Lemma 9 concerns an estimate for the mean switchover time given a certain number of message arrivals while switching.

LEMMA 9: *There is a positive constant α such that*

$$\mathbb{E}[R_2 \mid B_j] \le \alpha \left(j^2 + 1\right)$$

for all $j \in \mathbb{N}$ with the notation from (16).

SKETCH OF PROOF: We recall

$$\mathbb{E}[R_2 \mid B_j] = \frac{\mathbb{E}_R\left[R^{j+1} e^{-R}\right]}{\lambda_1 \mathbb{E}_R\left[R^j e^{-R}\right]}$$

for $R \coloneqq \lambda_1 R_2$ from (17) and we introduce the random variable X by

$$\mathbb{E}\left[f(X)\right] \coloneqq \frac{\mathbb{E}_R\left[f(R) e^{-R}\right]}{\mathbb{E}_R\left[e^{-R}\right]}, \quad f \in C_b.$$

Then, X has some finite exponential moment and one can show by elementary calculations that there is an $\alpha > 0$ such that

$$\frac{\mathbb{E}\left[X^{j+1}\right]}{\mathbb{E}\left[X^j\right]} \le \lambda_1 \alpha \left(j^2 + 1\right)$$

for all $j \in \mathbb{N}$. This finishes the proof. ■

The next Lemma 10 captures the fact that if there may be an additional waiting time due to a larger wait-and-see parameter $\tilde{T}_1 \ge T_1$, rather more messages arrive per cycle. Therefore, the probability of finding an empty queue upon arrival at station 1 becomes smaller.

LEMMA 10: *Consider a polling model with $N = 2$ stations, Strategy II and $T_2 = 0$. Given a $\bar{T}_1 > 0$, we have*

$$\pi_{\inf}(\bar{T}_1) \coloneqq \inf_{T_1 \in [0, \bar{T}_1]} \pi_0^{(1)}(T_1) > 0.$$

SKETCH OF PROOF: We can construct an appropriate coupling of two processes representing the polling models with wait-and-see parameters T_1 and $\tilde{T}_1$ for $0 \le T_1 \le \tilde{T}_1$. Due to the

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

construction, the queue length at station 1 upon exit from station 2 is always larger for the process with $\tilde{T}_1$ instead of T_1. Combining this observation and the ergodic theorem for Markov chains, we obtain

$$\nu_0^{(2)}(T_1) \geq \nu_0^{(2)}(\tilde{T}_1).$$

This inequality is equivalent to

$$\pi_0^{(1)}(T_1) \geq \pi_0^{(1)}(\tilde{T}_1)$$

due to (12). Then, we get $\pi_{\inf}(\bar{T}_1) = \pi_0^{(1)}(\bar{T}_1)$. ■

We make use of Theorems 1 and 2 to prove whether it is worth waiting. For the purpose of comparison, we recall the formula

$$\bar{D}^{\text{exh}} = \frac{\sum_{i=1}^{2} \lambda_i b_i^{(2)}}{2(1-\rho_0)} + \frac{r_0 \rho_1 \rho_2}{\rho_0(1-\rho_0)} + \frac{r_0^{(2)}}{2r_0}$$

for the mean average queueing delay of a message in a polling model with the exhaustive strategy from (2) by setting $f_1 = f_2 = 0$. Thus, we can rearrange the formula for the delay into $\bar{D} = \bar{D}^{\text{exh}} + \Delta\bar{D}$ with

$$\begin{aligned} \Delta\bar{D} := &- \frac{r_0^{(2)}}{2r_0} + \frac{r_0^{(2)}}{2(r_0+f_0)} \\ &+ \frac{\rho_2 f_1}{\rho_0(r_0+f_0)}(r_1 + \tilde{r}_2 + w_1) \\ &+ \frac{\rho_1 f_2}{\rho_0(r_0+f_0)}(\tilde{r}_1 + r_2 + w_2), \end{aligned} \tag{19}$$

where $\tilde{r}_i = r_i$ holds in the case of Strategy III (cf. Remark 7).

3.3.2. Proof of Theorem 3. Due to $T_2 = 0$, we have $f_2 = 0$ and the last line in (19) vanishes. It is worth waiting at station 1 if and only if there is a positive parameter of the wait-and-see strategy such that $\Delta\bar{D} < 0$. Since the expected waiting time at station 1 equals the total expected waiting time per cycle ($f_1 = f_0$), we rearrange inequality $\Delta\bar{D} < 0$ into

$$\frac{1}{r_0+f_1}\left[\frac{r_0^{(2)}}{2} + \frac{\rho_2}{\rho_0} f_1 (r_1 + \tilde{r}_2 + w_1)\right] < \frac{r_0^{(2)}}{2r_0}$$

whose validity is equivalent to

$$\left[-\frac{r_0^{(2)}}{2r_0} + \frac{\rho_2}{\rho_0}(r_1 + \tilde{r}_2 + w_1)\right] f_1 < 0. \tag{20}$$

We recall that w_i and f_i are non-negative quantities. Moreover, we observe that $f_i > 0$ holds for all $T_i > 0$. This can be argued by using the expected sojourn times for Strategy III and by using the steady-state probabilities for Strategies II and IV.

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

Strategy III. Note that we have $r_1 + \tilde{r}_2 = r_0$ according to Remark 7. For all $T_1 > 0$, we see from (10) that w_1 is greater than the expected length of the first busy period at station 1, that is, there is a function $\Delta_1(T_1) > 0$ such that

$$w_1 = (r_0 + c_2)\frac{\rho_1}{1-\rho_1} + \Delta_1(T_1).$$

We insert this representation of w_1 into (20), make use of (9) and obtain that (20) is equivalent to

$$-\frac{r_0^{(2)}}{2r_0} + \frac{\rho_2}{\rho_0}\left(\frac{1-\rho_2}{1-\rho_0}r_0 + \frac{\rho_1\rho_2}{1-\rho_0}\left(T_1 + \frac{q_1(T_1)b_1}{1-\rho_1}\right) + \Delta_1(T_1)\right) < 0. \tag{21}$$

Because of the property that both functions $\Delta_1(T_1)$ and $q_1(T_1)$ converge to zero for $T_1 \to 0$, we find the sufficient condition

$$\frac{r_0^{(2)}}{2r_0^2} - \frac{\rho_2(1-\rho_2)}{\rho_0(1-\rho_0)} > 0 \tag{22}$$

for "it is worth waiting at station 1". In order to establish the necessity of this condition, we argue in the following way: If we assume that (22) does not hold, inequality (21) is not satisfied for all $T_1 > 0$ because $\Delta_1(T_1)$ and $q_1(T_1)$ are non-negative, and we see that it is not worth waiting at station 1.

Strategy IV. The difference to Strategy III is the fact that w_1 does not have to be greater than the expected length of the first busy period at station 1. We just focus on

$$-\frac{r_0^{(2)}}{2r_0} + \frac{\rho_2}{\rho_0}(r_1 + \tilde{r}_2 + w_1) < 0 \tag{23}$$

from (20) and observe the property $w_1 \le T_1$ because a waiting period ends at the latest when the timer expires. In the same manner as above, we get the necessary and sufficient condition

$$\frac{r_0^{(2)}}{2r_0\left(r_1 + \tilde{r}_2^{\mathrm{IV}}\right)} - \frac{\rho_2}{\rho_0} > 0$$

for "it is worth waiting at station 1" with $\tilde{r}_2^{\mathrm{IV}}$ given by (18). In the case of deterministic switchover times, we just replace $r_0^{(2)}$ by r_0^2 and $\tilde{r}_2^{\mathrm{IV}}$ by r_2.

Strategy II. We focus again on (23) as with Strategy IV, and $w_1 \le T_1$ holds since waiting periods can only happen within the minimum sojourn time T_1. Differently from Strategy IV, the conditional mean switchover time $\tilde{r}_2^{\mathrm{II}}$ depends on the parameter T_1.

First, we prove that it is worth waiting with Strategy IV if it is worth waiting with Strategy II. Therefore, we assume that there is a $T_1 > 0$ such that (23) holds for Strategy II. We have to conclude that (3) is satisfied which can be easily seen if we have $\tilde{r}_2^{\mathrm{IV}} \le \tilde{r}_2^{\mathrm{II}}(T_1)$

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

for all $T_1 > 0$. We continue with proving this inequality. We recall

$$\tilde{r}_2^{\mathrm{II}} = \sum_{k=0}^{\infty} p_k^{(1)} \sum_{j=0}^{k} \frac{\mathbb{P}(A_{k-j})\mathbb{P}(B_j)}{\mathbb{P}(C_k)} \mathbb{E}[R_2 \mid B_j]$$

from (14) and (15), and

$$\tilde{r}_2^{\mathrm{IV}} = \mathbb{E}[R_2 \mid B_0]$$

from (18). We use the representation of $\mathbb{E}[R_2 \mid B_j]$ from (17) and the Cauchy–Schwarz inequality to get

$$\mathbb{E}[R_2 \mid B_j] \le \mathbb{E}[R_2 \mid B_{j+1}]$$

for all $j \in \mathbb{N}_0$. This property suffices in order to conclude $\tilde{r}_2^{\mathrm{IV}} \le \tilde{r}_2^{\mathrm{II}}(T_1)$ for all $T_1 > 0$.

Next, we have to prove that it is worth waiting with Strategy II if it is worth waiting with Strategy IV. Let (3) be satisfied, that is, there is a $T_1^{\mathrm{IV}} > 0$ such that (23) holds for $\tilde{r}_2^{\mathrm{IV}}$ and $w_1^{\mathrm{IV}}(T_1^{\mathrm{IV}})$. We are done if there is a $T_1 > 0$ such that

$$\tilde{r}_2^{\mathrm{II}}(T_1) + w_1^{\mathrm{II}}(T_1) \le \tilde{r}_2^{\mathrm{IV}} + w_1^{\mathrm{IV}}(T_1^{\mathrm{IV}})$$

because (23) is the criterion for "it is worth waiting with Strategy II" as well. We observe

$$\begin{aligned}\tilde{r}_2^{\mathrm{II}} &= p_0^{(1)}\mathbb{E}[R_2 \mid B_0] + \sum_{k=1}^{\infty} p_k^{(1)} \sum_{j=0}^{k} \frac{\mathbb{P}(A_{k-j})\mathbb{P}(B_j)}{\mathbb{P}(C_k)} \mathbb{E}[R_2 \mid B_j] \\ &\le \mathbb{E}[R_2 \mid B_0] + \sum_{k=1}^{\infty} p_k^{(1)}\mathbb{E}[R_2 \mid B_k]\end{aligned}$$

and define

$$\varepsilon := \frac{w_1^{\mathrm{IV}}(T_1^{\mathrm{IV}})}{2}.$$

Due to $\tilde{r}_2^{\mathrm{IV}} = \mathbb{E}[R_2 \mid B_0]$ and $w_1^{\mathrm{II}}(T_1) \le T_1$, it suffices to show that there is a positive $T_1 < \varepsilon$ such that

$$\sum_{k=1}^{\infty} p_k^{(1)}\mathbb{E}[R_2 \mid B_k] < \varepsilon.$$

We recall

$$p_k^{(1)} = \frac{\pi_k^{(1)} \int_0^{T_1} P_{k,0}(x)\,dx}{\sum_{l=0}^{\infty} \pi_l^{(1)} \int_0^{T_1} P_{l,0}(x)\,dx}$$

from (13). First, we estimate the quantity $\int_0^{T_1} P_{k,0}(x)\,dx$ that is the expected length of the total waiting time during the stay at station 1 given that there are k messages waiting upon

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

arrival. We get

$$\begin{aligned}\int_0^{T_1} P_{0,0}(x)\,dx &\geq T_1\,\mathbb{P}(\text{no message arrives at station 1 within the time } T_1)\\ &= T_1 e^{-\lambda_1 T_1}\end{aligned}$$

and

$$\begin{aligned}\int_0^{T_1} P_{k,0}(x)\,dx &\leq T_1\,\mathbb{P}(\text{the length of the first busy period} \leq T_1)\\ &\leq T_1\,\mathbb{P}(\text{the sum of } k \text{ independent service times} \leq T_1)\\ &\leq T_1\left(1 - e^{-\mu_1 T_1}\sum_{j=0}^{k-1}\frac{(\mu_1 T_1)^j}{j!}\right)\\ &= T_1 e^{-\mu_1 T_1}\left(e^{\mu_1 T_1} - \sum_{j=0}^{k-1}\frac{(\mu_1 T_1)^j}{j!}\right)\\ &= T_1 e^{-\mu_1 T_1}\sum_{j=k}^{\infty}\frac{(\mu_1 T_1)^j}{j!}\\ &= T_1 e^{-\mu_1 T_1}(\mu_1 T_1)^k\sum_{j=0}^{\infty}\frac{(\mu_1 T_1)^j}{(j+k)\cdots(j+1)j!}\\ &\leq T_1(\mu_1 T_1)^k\end{aligned}$$

for all $k \in \mathbb{N}$ where we use the Erlang(k, μ_1) distribution function in the third line. Now, we can bound $p_k^{(1)}$ for all $k \in \mathbb{N}$ from above by

$$p_k^{(1)} \leq \frac{T_1(\mu_1 T_1)^k}{\pi_0^{(1)} T_1 e^{-\lambda_1 T_1}} = \frac{e^{\lambda_1 T_1}}{\pi_0^{(1)}}(\mu_1 T_1)^k.$$

Using Lemmas 9 and 10 with $\bar{T}_1 := 1/\mu_1$ in the first two lines and using limits of geometric series, we obtain for $T_1 \in (0, \bar{T}_1)$ with $q := \mu_1 T_1 < 1$

$$\begin{aligned}\sum_{k=1}^{\infty} p_k^{(1)}\mathbb{E}[R_2 \mid B_k] &\leq \sum_{k=1}^{\infty}\frac{e^{\lambda_1 T_1}}{\pi_0^{(1)}}(\mu_1 T_1)^k\alpha\left(k^2+1\right)\\ &\leq \frac{\alpha e^{\frac{\lambda_1}{\mu_1}}}{\pi_{\inf}(\bar{T}_1)}\left(\sum_{k=1}^{\infty}k^2 q^k + \sum_{k=1}^{\infty}q^k\right)\\ &= \frac{\alpha e^{\frac{\lambda_1}{\mu_1}}}{\pi_{\inf}(\bar{T}_1)}\left(\frac{q(1+q)}{(1-q)^3} + \frac{q}{1-q}\right).\end{aligned}$$

Finally, we are done because the term in the last line converges to zero for $T_1 \to 0$.

3.3.3. Proof of Theorem 5. We focus on inequality $\Delta\bar{D} < 0$ which can be rearranged into

$$\frac{1}{r_0 + f_0}\left[\frac{r_0^{(2)}}{2} + \frac{\rho_2}{\rho_0}f_1\left(r_1 + \tilde{r}_2 + w_1\right) + \frac{\rho_1}{\rho_0}f_2\left(\tilde{r}_1 + r_2 + w_2\right)\right] < \frac{r_0^{(2)}}{2r_0}.$$

https://doi.org/10.1017/S0269964817000419 Published online by Cambridge University Press

Strategy III. We can proceed in an analogous manner as in the proof of Theorem 3. Using the symmetry $\rho_1 = \rho_2$, we obtain the necessary and sufficient condition

$$\frac{r_0^{(2)}}{r_0^2} - \frac{1-\rho_1}{1-\rho_0} > 0$$

for "it is worth waiting" at both stations with $T_1 = T_2 > 0$.

Strategies II and IV. In addition to the procedure in the proofs above, we have to extend Lemma 10 by setting $T_1 = T_2 > 0$. Then, for a totally symmetric polling model we get the necessary and sufficient condition

$$\frac{r_0^{(2)}}{r_0\left(r_1 + \tilde{r}_2^{\mathrm{IV}}\right)} > 1 \tag{24}$$

for "it is worth waiting" at both stations in the same way. A short calculation shows that $\tilde{r}_2^{\mathrm{IV}} \leq r_2$ holds. Therefore, we can conclude that (24) is satisfied if and only if the switchover times are non-deterministic.

3.3.4. Proof of Corollary 6. We just have to set the system parameters such that the condition (inequality) in Theorem 3 or 5 is fulfilled for Strategies II and IV but not for Strategy III.

Acknowledgment

The work of S. Schwinn is supported by the Excellence Initiative of the German Federal and State Governments via the Graduate School of Computational Engineering at Technische Universität Darmstadt.

We would like to thank the referee for the careful review of our manuscript and for many comments that helped to improve the presentation of the paper.

References

1. Afanassieva, L.G., Delcoigne, F., & Fayolle, G. (1997). On polling systems where servers wait for customers. *Markov Processes and Related Fields* 3(4): 527–545.
2. Al Hanbali, A., de Haan, R., Boucherie, R.J., & van Ommeren, J.-K. (2012). Time-limited polling systems with batch arrivals and phase-type service times. *Annals of Operations Research* 198(1): 57–82.
3. Asmussen, S. (1987). *Applied probability and queues*. Wiley series in probability and mathematical statistics. Chichester: Wiley.
4. Aurzada, F., Beck, S., & Scheutzow, M. (2012). Wait-and-see strategies in polling models. *Probability in the Engineering and Informational Sciences* 26(1): 17–42.
5. Boxma, O.J. & Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability* 24(4): 949–964.
6. Boxma, O.J., Schlegel, S., & Yechiali, U. (2002). Two-queue polling models with a patient server. *Annals of Operations Research* 112(1): 101–121.
7. Cooper, R.B., Niu, S.-C., & Srinivasan, M.M. (1998). When does forced idle time improve performance in polling models? *Management Science* 44(8): 1079–1086.
8. de Haan, R. (2009). *Queueing models for mobile ad hoc networks*. Ph.D. thesis, University of Twente, Enschede.
9. de Haan, R., Boucherie, R.J., & van Ommeren, J.-K. (2009). A polling model with an autonomous server. *Queueing Systems* 62(3): 279–308.
10. de Souza e Silva, E., Gail, H.R., & Muntz, R.R. (1995). Polling systems with server timeouts and their application to token passing networks. *IEEE/ACM Transactions on Networking* 3(5): 560–575.
11. Eliazar, I. & Yechiali, U. (1998). Polling under the randomly timed gated regime. *Communications in Statistics. Stochastic Models* 14(1–2): 79–93.

12. Frigui, I. & Alfa, A.-S. (1998). Analysis of a time-limited polling system. *Computer Communications* 21(6): 558–571.
13. Kleinrock, L. (1975). *Queueing systems, Volume I: Theory*. New York: Wiley.
14. Leung, K.K. (1994). Cyclic-service systems with nonpreemptive, time-limited service. *IEEE Transactions on Communications* 42(8): 2521–2524.
15. Li, J.Z. (2009). Two-queue polling model with a timer and a randomly-timed gated mechanism. *Journal of Mathematical Research and Exposition* 29(4): 721–729.
16. 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.
17. Samaddar, S. & Whalen, T. (2008). Improving performance in cyclic production systems by using forced variable idle setup time. *Manufacturing & Service Operations Management* 10(2): 173–180.
18. Sarkar, D. & Zangwill, W.I. (1991). Variance effects in cyclic production systems. *Management Science* 37(4): 444–453.
19. Takagi, H. (1986). *Analysis of polling systems*. Cambridge: The MIT Press.
20. Xie, J., Fischer, M.J., & Harris, C.M. (1997). Workload and waiting time in a fixed-time loop system. *Computers & Operations Research* 24(8): 789–803.