TIME-LIMITED POLLING SYSTEMS WITH BATCH ARRIVALS AND PHASE-TYPE SERVICE TIMES

AHMAD AL HANBALI, EURANDOM, P.O. BOX 513, 5600 MB EINDHOVEN, THE NETHERLANDS,
ALHANBALI@EURANDOM.TUE.NL

ROLAND DE HAAN, RICHARD J. BOUCHERIE, AND JAN-KEES VAN OMMEREN ,
UNIVERSITY OF TWENTE, P.O. BOX 217, 7500 AE ENSCHEDE, THE NETHERLANDS
{R.DEHAAN,R.J.BOUCHERIE,J.C.W.VANOMMEREN}@UTWENTE.NL

Abstract. In this paper, we will develop a general framework to analyze polling systems with either the autonomous-server or the time-limited service discipline. We consider Poisson batch arrivals and phase-type service times. It is known that these disciplines do not satisfy the well-known branching property in polling system. Therefore, hardly any exact results exist in the literature. Our strategy is to apply an iterative scheme that is based on relating in closed-form the joint queue-length at the beginning and the end of a server visit to a queue. These kernel relations are derived using the theory of absorbing Markov chains.

Keywords: Absorbing Markov chains. Matrix analytic solution. Polling system. Autonomous server discipline. Time limited discipline. Poisson batch arrivals. Phase-type service times. Iterative scheme. Performance analysis.

1. Introduction

Polling systems have been extensively studied in the last years due to their vast area of applications in production and telecommunication systems [15, 18]. They have demonstrated to offer an adequate modeling framework to analyze systems in which a set of entities need certain service from a single resource. These entities are located at different positions in the system awaiting their turn to receive service.

In queueing theory, a polling system is equivalent to a set of queues with exogenous job arrivals all requiring service from a single server. The server serves each queue according to a specific service discipline and after serving a queue he will move to a next queue. A tractable analysis of a polling system is possible if the system satisfies the so-called branching property [17]. This property states that each job present at a queue at the arrival instant of the server will be replaced in an independent and identically distributed manner by a random number of jobs during the course of the server's visit. For disciplines not satisfying this property hardly any exact results are known.

The two most well-known disciplines that satisfy the branching property are the exhaustive and gated discipline. Exhaustive means that the server continues servicing a queue until it becomes empty. At this instant the server moves to the next queue in his schedule. Gated means that the server only serves the jobs present in the queue at its arrival.

The drawback of the exhaustive and gated disciplines is that the server is controlled by the presence of jobs at Q_i. To reduce this control on the server, other type of service disciplines were introduced such as the time-limited or the k-limited discipline. According to the time-limited discipline, the server continues servicing a queue for a certain time period or until the queue becomes empty, whichever occurs first. Under the k-limited discipline, the server continues servicing a queue until k jobs are served or the queue becomes empty, whichever occurs first. Another discipline, evaluated more recently in the literature and closely related to the time-limited discipline, is the so-called autonomous-server discipline [1, 8], where the server stays at a queue for a certain period of time, even if the queue becomes empty. This discipline may also be seen as the non-exhaustive time-limited discipline. We should emphasize that these latter disciplines do not satisfy the branching property and thus hardly any closed-form results are known for the queue-length distribution under these disciplines.

To circumvent this difficulty, researchers resort to numerical methods using for instance iterative solution techniques or the power series algorithm. The power series algorithm [4, 5] aims at solving the global balance equations. To this end, the state probabilities are written as a power series and via a complex computation scheme the coefficients of these series, and thus the queue-length probabilities, are obtained. The iterative techniques [13, 14] exploit the relations between the joint queue-length distributions at specific instants, viz., the start of a server visit and the end of a server visit. The relation between the queue length at the start and end of a visit to a queue is established via recursively expressing the queue length at a job departure instant in terms of the queue length at the previous departure instant of a job. The complementary relation, between the queue length at the end of a visit to a queue and a start of visit to a next queue, can easily be established via the switch-over time. Starting with an initial distribution, the stationary queue-length distribution is then obtained by means of iteration. For the k-limited discipline, the authors in [20] proposed an iterative approximation that is based on a matrix geometric method. Although these methods offer a way to numerically solve intrinsically hard systems, their solution provides little fundamental insight.

Under the assumption of exponential service times, we derived in [2] a direct and more insightful relation between the joint number of jobs at the beginning and end of a server visit to a queue for the autonomous-server, the time-limited, and the k-limited discipline. This is done using a matrix analytic approach. In the same paper, we also re-derived a result of [21] for the exhaustive time-limited discipline for the special case of exponential service times. The latter article studied the exhaustive time-limited discipline for preemptive service [21]. Observing that upon successful service completion at a queue the busy period in fact regenerates, the authors could obtain a closed-form relation between the joint queue length at the end and the beginning of a server visit. In [7] all these results were extended by including routing of jobs between the different queues. This is done by constructing Markov chains at specific embedded epochs and subsequently relating the state space at these epochs.

In this paper, we develop a framework to analyze the autonomous server and the time-limited polling systems with Poisson batch arrivals and phase-type service times. Our framework incorporates an iterative solution method which enhances the method introduced in [13]. More specifically, contrary to that approach, we will establish a direct relation between the joint number of jobs at the beginning and end of a server visit to a queue without conditioning on any intermediate events that occur during a visit. To this end, we use the theory of absorbing Markov chains (AMC) [11, 16]. We construct an AMC whose transient states represent the states of the polling system. The event of the server leaving a queue is modeled as an absorbing event. We will set the initial state of the AMC to the joint number of jobs at the beginning of a service period of a queue. Therefore, to find the joint number of jobs at the end of a service period, it is sufficient to keep track of the state from which the transition to the absorption state occurs. The probability of the latter event is eventually determined by first ordering the states in a careful way and consequently exploiting the structures that arise in the generator matrix of the AMC. Following this approach, we relate in closed-form the joint queue-length probability generating functions (p.g.f.) at the end of a visit period to a queue to the joint queue-length p.g.f. at the beginning of this visit period. The major part of this paper is devoted to deriving these kernel relations for the above-mentioned two disciplines: autonomous-server and time-limited. Once these relations are obtained, the joint queue-length distribution at server departure instants is readily obtained via a simple iterative scheme.

Although we have developed our framework for the case of autonomous-server and time-limited systems, our framework is generally applicable to analyze other branching and non-branching type polling systems. The key step is the correct ordering of the states that allows us to invoke the theory of absorbing Markov chains in order to relate in closed-form the joint number of jobs in the system at the beginning and end of a server visit to a queue.

The paper is organized as follows. In Section 2 we give a detailed description of the model and the assumptions. Section 3 analyzes the autonomous-server discipline. In Section 4 we study the time-limited discipline. In Section 5 we describe the iterative scheme that is important to compute the joint queue-length distribution. Finally, in Section 6, we conclude the paper and give some research directions.

2. Model

We consider a single-server polling model consisting of M first-in-first-out (FIFO) systems with unlimited queue, Q_i, $i = 1, \dots, M$. Jobs arrive to Q_i in batches according to a Poisson process of rate λ_i. The sequence of batch sizes consists of independent and identically distributed random variables, which are independent of inter-arrival times. Let us denote D_i the batch size at Q_i with probability mass function $D_i(\cdot)$ and probability generating function $\hat{D}_i(z)$, $|z| \leq 1$. We assume that $D_i \geq 1$ for $i = 1, \dots, M$. The service time of a job at Q_i is denoted by B_i. B_i is a phase-type random variable with distribution function $B_i(\cdot)$ with mean b_i and h_i phases. That is, B_i is a mixture of h_i exponential random variables. We assume that the service requirements are independent and identically distributed random variables and they are independent of the batch size and inter-arrival time.

A phase-type distribution can be represented by an initial distribution vector π, a transient generator $\mathbf{T}$, and an absorption rate vector T^o, i.e., $\mathbf{T}^{-1}T^0 = -e^T$, where e^T is a column vector with all entries equal to one. For more details we refer, e.g., to [16, p. 44]. Then, it is well-known that the Laplace-Stieltjes transform (LST) of the service times at Q_i, B_i, can written as follows

$$\tilde{B}_i(s) = \pi_i(s\mathbf{I} - \mathbf{T}_i)^{-1}T_i^o, \quad \mathrm{Re}(s) \geq 0. \tag{1}$$

For later use, we need to introduce the LST of residual (phase-type) service times.

Lemma 1. *The LST of the residual service times at Q_i is given by*

$$\tilde{B}_i^*(s) = \frac{1}{b_i}\pi_i(s\mathbf{I} - \mathbf{T}_i)^{-1}e^T, \quad \text{Re}(s) \geq 0. \tag{2}$$

Proof. The LST of the residual service times reads

$$\begin{aligned}\tilde{B}_i^*(s) &= \frac{1}{b_i s}(1 - \tilde{B}_i(s)) \\ &= -\frac{1}{b_i}\pi_i\mathbf{T}_i^{-1}(s\mathbf{T}_i^{-1} - \mathbf{I})^{-1}\mathbf{T}_i^{-1}T_i^o \\ &= \frac{1}{b_i}\pi_i(s\mathbf{I} - \mathbf{T}_i)^{-1}e^T.\end{aligned}$$

□

We let $N_i(t)$ denote the number of jobs in Q_i, $i = 1, \ldots, M$, at time $t \geq 0$ and it is assumed that $N_i(0) = 0$, $i = 1, \ldots, M$. The server visits the queues in a cyclic fashion. After a visit to Q_i, the server incurs a switch-over time C^i from Q_i to Q_{i+1}. We assume that C^i is independent of the service requirement and follows a general distribution $C^i(\cdot)$ with mean c^i, where at least one $c^i > 0$. The service discipline at each queue is either autonomous-server or time-limited. Under the autonomous-server discipline, the server remains at location Q_i an exponentially distributed time with rate α_i before it migrates to the next queue in the cycle. Under the time-limited discipline, the server departs from Q_i when it becomes empty or when a timer of exponentially duration with rate α_i has expired, whichever occurs first.

It is assumed that the queues of the polling system are stable. In the following lemmas we shall report the stability condition for both the autonomous-server and the time-limited systems. The proofs of these lemmas are straightforward extensions to those of Theorems 3.1 and 3.2 in [7, Chap. 3].

Lemma 2 (Autonomous-server discipline)**.**

$$\textit{System is stable} \Longleftrightarrow \rho_i < \kappa_i, \quad i = 1, \ldots, M,$$

where

$$\rho_i = \lambda_i \mathbb{E}[D_i] \cdot \frac{1 - \tilde{B}_i(\alpha_i)}{\alpha_i \tilde{B}_i(\alpha_i)}, \quad \kappa_i = \frac{1/\alpha_i}{\sum_{j=1}^M 1/\alpha_j + c_j}.$$

We note that $\big(1 - \tilde{B}_i(\alpha_i)\big)/(\alpha_i \tilde{B}_i(\alpha_i))$ is the LST of the *effective service times* of a job in Q_i which includes the work lost due to service preemptions. κ_i is the availability fraction of the server at Q_i.

Lemma 3 (Time-limited discipline)**.**

$$\textit{System is stable} \Longleftrightarrow \rho + \max_{i=1,\ldots,M}\left(\frac{\lambda_i \mathbb{E}[D_i]}{\mathbb{E}[G_i^*]}\right) \cdot c_t < 1,$$

where

$$\rho = \sum_{j=1}^M \frac{\lambda_i \mathbb{E}[D_i]\big(1 - \tilde{B}_i(\alpha_i)\big)}{\alpha_i \tilde{B}_i(\alpha_i)}, \quad \mathbb{E}[G_i^*] = \frac{\tilde{B}_i(\alpha_i)}{1 - \tilde{B}_i(\alpha_i)}, \quad c_t = \sum_{j=1}^M c_j.$$

We note that ρ represents the total offered load to the system and $\mathbb{E}[G_i^*]$ the mean number of served jobs at Q_i during a cycle when Q_i is saturated.

In case the server is active at the end of a server visit, which may happen under the autonomous-server and time-limited disciplines, then the service will be preempted. At the beginning of the next visit of the server, the service time will be re-sampled according to $B_i(\cdot)$. This discipline is commonly referred to as *preemptive-repeat-random*.

A word on notation. Given a random variable X, $X(t)$ will denote its distribution function. We use $\mathbf{I}$ to denote an identity matrix of appropriate size and use $\otimes$ as the Kronecker product operator defined as follows. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices and $a(i,j)$ and $b(i,j)$ denote the (i,j)-entries of $\mathbf{A}$ and $\mathbf{B}$ respectively then $\mathbf{A}\otimes\mathbf{B}$ is a block matrix where the (i,j)-block is equal to $a(i,j)\mathbf{B}$. We use e to denote a row vector of appropriate size with entries equal to one and e_i to denote a row vector of appropriate size with the i-th entry equal to one and the other elements equal to zero. Finally, v^T will denote the transpose of vector v.

3. Autonomous-server discipline

In this section, we will relate the joint queue-length probabilities at the beginning and end of a server visit to a queue for the autonomous-server discipline. Under the autonomous-server discipline, the server remains at location Q_i an exponentially distributed time with rate α_i before it migrates to the next queue in the cycle. It is stressed that even when Q_i becomes empty, the server will remain at this queue.

Without loss of generality let us consider a server visit to Q_1. The number of jobs at the various queues at the beginning of a server visit to Q_1 is denoted by $\mathbf{N}_1^b := (N_{11}^b,\dots,N_{M1}^b)$; let $\mathbf{N}_1^e := (N_{11}^e,\dots,N_{M1}^e)$ denote the queue lengths at the end of such a visit. We assume that the p.g.f. of the steady-state queue-length at service's beginning instant at Q_1, denoted by $\beta_1^A(\mathbf{z}) = \mathbb{E}\left[\mathbf{z}^{\mathbf{N}_1^b}\right]$, is known, where $\mathbf{z} := (z_1,\dots,z_M)$ and $|z_i|\le 1$ for $i=1,\dots,M$. The aim is to derive the p.g.f. of the steady-state queue-length at service visit's end at Q_1, denoted by $\gamma_1^A(\mathbf{z}) = \mathbb{E}\left[\mathbf{z}^{\mathbf{N}_1^e}\right]$.

Let $\mathbf{N}(t) := \big(PH_1(t), N_1(t),\dots,N_M(t)\big)$ denote the $(M+1)$-dimensional, continuous-time Markov chain with discrete state-space $\xi_A = \{0,1,\dots,h_1\}\times\{0,1,\dots\}^M \cup \{a\}$, where $N_m(t)$, $m = 1\dots,M$, represents the number of jobs in Q_m and $PH_1(t)$ the phase of the job in service at Q_1 at time t. State $\{a\}$ is absorbing. We refer to this absorbing Markov chain by $\mathbf{AMC}_A$. The absorption of $\mathbf{AMC}_A$ occurs when the server leaves Q_1 which happens with rate α_1. Moreover, the initial state of $\mathbf{AMC}_A$ at $t=0$ is set to the system state at server's arrival to Q_1, i.e., $\mathbf{N}_1^b = (i_1,\dots,i_M)$. Therefore, the probability that the absorption of $\mathbf{AMC}_A$ occurs from state $(j_1,\dots,j_M)$ equals $\mathbb{P}\big(\mathbf{N}_1^e = (j_1,\dots,j_M) \bigm| \mathbf{N}_1^b = (i_1,\dots,i_M)\big)$.

We derive now $\mathbb{P}\big(\mathbf{N}_1^e = (j_1,\dots,j_M)|\ \mathbf{N}_1^b = (i_1,\dots,i_M)\big)$. During a server visit to Q_1, the number of jobs at Q_m, $m=2,\dots,M$, may only increase. Therefore, $\mathbb{P}\big(\mathbf{N}_1^e = (j_1,\dots,j_M) \mid \mathbf{N}_1^b = (i_1,\dots,i_M)\big) = 0$ for $j_l < i_l$, $l=2,\dots,M$. For sake of clarity, we shall show first in detail the structure of $\mathbf{AMC}_A$ in the case of 3 queues, i.e. for $M=3$, and the procedure of the proof of the desired result before considering the general case.

Case M=3. Let us consider the transient states of $\mathbf{AMC}_A$, i.e., $(ph_1, n_1, n_2, n_3) \in$

$\xi_A \setminus \{a\}$. We recall that we consider a server visit to Q_1. The number of jobs at Q_2 and Q_3 may only increase during a server visit to Q_1, while the number of jobs at Q_1 may increase or decrease. To take advantage of this property, we will order the transient states of the $\mathbf{AMC}_A$ as follows: $(0,0,0,0),(1,0,0,0),\ldots,(0,1,0,0),(1,1,0,0)$, $\ldots,(0,0,1,0),(1,0,1,0),\ldots,(0,0,0,1),(1,0,0,1),\ldots$, i.e., lexicographically ordered first according to n_3, then n_2, n_1, and finally according to ph_1. This ordering induces that the generator matrix of the transitions between the transient states of $\mathbf{AMC}_A$ for $M=3$, denoted by $\mathbf{Q}_3$, is an infinite upper-triangular block matrix with diagonal blocks equal to $\mathbf{A}_3$ and i-th upper-diagonal blocks equal $\lambda_3 D_3(i)\mathbf{I}$, i.e.,

$$\mathbf{Q}_3 = \begin{pmatrix} \mathbf{A}_3 & \lambda_3 D_3(1)\mathbf{I} & \lambda_3 D_3(2)\mathbf{I} & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_3 & \lambda_3 D_3(1)\mathbf{I} & \lambda_3 D_3(2)\mathbf{I} & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3}$$

We note that $\mathbf{A}_3$ denotes the generator matrix of the transitions which do not induce any modification in the number of jobs at Q_3. Moreover, $\lambda_3 D_3(i)\mathbf{I}$ denotes the transition rate matrix between the transient states (ph_1, n_1, n_2, n_3) and (ph_1, n_1, n_2, n_3+i), i.e., the transitions that represent an arrival of a batch of size i to Q_3. The block matrix $\mathbf{A}_3$ is also an infinite upper-triangular block matrix with diagonal blocks equal to $\mathbf{A}_2$, and i-th upper-diagonal blocks equal $\lambda_2 D_2(i)\mathbf{I}$, i.e.,

$$\mathbf{A}_3 = \begin{pmatrix} \mathbf{A}_2 & \lambda_2 D_2(1)\mathbf{I} & \lambda_2 D_2(2)\mathbf{I} & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_2 & \lambda_2 D_2(1)\mathbf{I} & \lambda_2 D_2(2)\mathbf{I} & \cdots & \cdots \\ \vdots\ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4}$$

where $\lambda_2 D_2(i)\mathbf{I}$ denotes the transition rate matrix between the states (ph_1, n_1, n_2, n_3) and (ph_1, n_1, n_2+i, n_3). $\mathbf{A}_2$ is the generator matrix of the transition between the states (ph_1, n_1, n_2, n_3) and (l, k, n_2, n_3) with $k \geq \max(n_1-1, 0)$ and $l \leq h_1$, the total number of phases in the service times. Observe that $\mathbf{A}_2$ equals the sum of the matrix $-(\lambda_2+\lambda_3+\alpha_1)\mathbf{I}$ and the generator matrix of an M^X/PH/1 queue with Poisson batch arrivals and phase-type service times. Let $\mathbf{A}_1$ denote the generator of an M^X/PH/1. It is readily seen that (see, e.g., [16, Chap. 3, Sec. 2])

$$\mathbf{A}_1 = \begin{pmatrix} -\lambda_1 & \lambda_1 D_1(1)\pi_1 & \lambda_1 D_1(2)\pi_1 & \cdots & \cdots & \cdots \\ T_1^o & \mathbf{T}_1 - \lambda_1\mathbf{I} & \lambda_1 D_1(1)\mathbf{I} & \lambda_1 D_1(2)\mathbf{I} & \cdots & \cdots \\ \mathbf{0} & T_1^o\pi_1 & \mathbf{T}_1 - \lambda_1\mathbf{I} & \lambda_1 D_1(1)\mathbf{I} & \lambda_1 D_1(2)\mathbf{I} & \cdots \\ \vdots\ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{5}$$

We recall that T_1^o is a column vector and π_1 is a row vector thus $T_1^o\pi_1$ is a matrix of rank one with (i,j)-entry representing the transition rate from state (i, n_1, n_2, n_3) to $(j, n_1 - 1, n_2, n_3)$.

Now, we compute $\mathbb{P}\big(\mathbf{N}_1^e = (j_1, j_2, j_3) \mid \mathbf{N}_1^b = (i_1, i_2, i_3)\big)$ as function of the inverse of $\mathbf{Q}_3$, $\mathbf{A}_3$ and $\mathbf{A}_2$ and later on we shall uncondition on N_{13}^e, then on N_{12}^e, and finally on N_{11}^e. We emphasize that since $\mathbf{Q}_3$, $\mathbf{A}_3$ and $\mathbf{A}_2$ are all sub-generators with the sum of their row elements strictly negative, these matrices are invertible. It shall become clear that in this paper we do not need to determine these inverse matrices in closed-form. For conveniance, we abbreviate the condition $\mathbf{N}_1^b = (i_1, i_2, i_3)$ to $\mathbf{N}_1^b$, e.g., $\mathbb{P}\big(\mathbf{N}_1^e = (j_1, j_2, j_3) \mid \mathbf{N}_1^b\big)$ denotes $\mathbb{P}\big(\mathbf{N}_1^e = (j_1, j_2, j_3) \mid \mathbf{N}_1^b = (i_1, i_2, i_3)\big)$.

From the theory of absorbing Markov chains, given that $\mathbf{AMC}_A$ starts in state $\mathbf{N}_1^b = (i_1, i_2, i_3)$, the probability that the transition to the absorption state $\{a\}$ occurs from state (j_1, j_2, j_3) reads (see, e.g., [10])

$$\mathbb{P}\big(\mathbf{N}_1^e = (j_1, j_2, j_3) \mid \mathbf{N}_1^b\big) = -\alpha_1 c_3 (\mathbf{Q}_3)^{-1} d_3, \tag{6}$$

where c_3 is the probability distribution vector of $\mathbf{AMC}_A$'s initial state which is given by

$$c_3 := e_{i_3} \otimes e_{i_2} \otimes e_{i_1} \otimes \pi_1,$$

and $\alpha_1 d_3$ is the transition rate vector to $\{a\}$ given that (j_1, j_2, j_3) is the last state visited before absorption with

$$d_3 := e_{j_3} \otimes e_{j_2} \otimes e_{j_1} \otimes e.$$

Note that the presence of π_1 in c_3 is due to the preemptive-repeat discipline, and e in d_3 is due to the un-conditioning on the phase of the service times in Q_1 when the server leaves the queue. We note that in [12] the absorption probability was introduced in terms of Palm measures and was applied on infinite state space absorbing Markov chains.

For later use, let us define the following row vectors:

$$c_2 := e_{i_2} \otimes e_{i_1} \otimes \pi_1, \quad d_2 := e_{j_2} \otimes e_{j_1} \otimes e,$$

$$c_1 := e_{i_1} \otimes \pi_1, \quad d_1 := e_{j_1} \otimes e.$$

We are now ready to formulate our first result.

Lemma 4. *The conditional generating function of the queue-length of Q_3 at the end of the server visit to Q_1 is given by*

$$\mathbb{E}\big[\, z_3^{N_{31}^e} \mathbf{1}_{\{N_{11}^e = j_1, N_{21}^e = j_2\}} \,\big|\, \mathbf{N}_1^b \big] \;=\; -\alpha_1 z_3^{i_3} c_2 \big(\lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_3}\big)^{-1} d_2^T. \tag{7}$$

Proof. Multiplying (6) by $z_3^{j_3}$ and summing these equations over j_3 we find that

$$\begin{aligned} \mathbb{E}\big[\, z_3^{N_{31}^e} \mathbf{1}_{\{N_{11}^e = j_1, N_{21}^e = j_2\}} \,\big|\, \mathbf{N}_1^b \big] &= -\alpha_1 c_3 (\mathbf{Q}_3)^{-1} \sum_{j_3 \ge i_3} z_3^{j_3} (e_{j_3} \otimes d_2)^T \\ &= -\alpha_1 c_3 (\mathbf{Q}_3)^{-1} \Big(\sum_{j_3 \ge i_3} z_3^{j_3} e_{j_3} \otimes d_2 \Big)^T \\ &= -\alpha_1 \Big(\sum_{j_3 \ge i_3} z_3^{j_3} u_3(j_3) \Big) d_2^T, \end{aligned} \tag{8}$$

where $\mathbf{u}_3 = \big(u_3(0), u_3(1), \dots\big) := c_3(\mathbf{Q}_3)^{-1}$. First, let us derive $\sum_{j_3 \ge i_3} z_3^{j_3} u_3(j_3)$. Note that $\mathbf{u}_3 \mathbf{Q}_3 = c_3$. Inserting $\mathbf{Q}_3$ given in (3) into the latter equation gives that

$$u_3(0)\mathbf{A_3} = \mathbf{0}, \tag{9}$$

$$\lambda_3 \sum_{l=0}^{n-1} D_3(n-l) u_3(l)\mathbf{I} + u_3(n)\mathbf{A_3} = \mathbf{1}_{\{n=i_3\}} c_2, \quad n \ge 1. \tag{10}$$

Note, since $\mathbf{A_3}$ is nonsingular, Eq. (9) yields that $u_3(0) = \mathbf{0}$, i.e., $u_3(0)$ is a vector of zeros. Inserting $u_3(0) = \mathbf{0}$ into (10) with $n = 1$ yields that $u_3(1) = \mathbf{0}$. Therefore, we deduce by an induction argument that $u_3(n) = \mathbf{0}$ for $n = 0, \ldots, i_3 - 1$. The latter system of equations now rewrites

$$u_3(i_3)\mathbf{A_3} = c_2, \tag{11}$$

$$\lambda_3 \sum_{l=i_3}^{n-1} D_3(n-l)u_3(l) + u_3(n)\mathbf{A_3} = \mathbf{0}, \quad n > i_3. \tag{12}$$

Multiplying (11) by $z_3^{i_3}$ and (12) by z_3^n and summing these equations over n we find that

$$\sum_{j_3 \geq i_3} z_3^{j_3} u_3(j_3) = z_3^{i_3} c_2 \big(\lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_3}\big)^{-1}. \tag{13}$$

Inserting (13) into (8) readily gives Lemma 4. □

Lemma 5. *The conditional generating function of the joint queue-length of Q_2 and Q_3 at the end of the server visit to Q_1 is given by*

$$\mathbb{E}\Big[\, z_2^{N^e_{21}} z_3^{N^e_{31}} \mathbf{1}_{\{N^e_{11}=j_1\}} \,\Big|\, \mathbf{N}^b_1\Big] = -\alpha_1 z_2^{i_2} z_3^{i_3} c_1 \big(\lambda_2 \hat{D}_2(z_2)\mathbf{I} + \lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_2}\big)^{-1} d_1^T. \tag{14}$$

Proof. Multiplying (7) by $z_2^{j_2}$ and summing over j_2 gives that

$$\begin{aligned}\mathbb{E}\Big[\, z_2^{N^e_{21}} z_3^{N^e_{31}} \mathbf{1}_{\{N^e_{11}=j_1\}} \,\Big|\, \mathbf{N}^b_1\Big] &= -\alpha_1 z_3^{i_3} c_2 \big(\lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_3}\big)^{-1} \Big(\sum_{j_2 \geq i_2} z_2^{j_2} e_{j_2} \otimes d_1\Big)^T \\ &= -\alpha_1 z_3^{i_3} \Big(\sum_{j_2 \geq i_2} z_2^{j_2} u_2(j_2)\Big) d_1^T, \end{aligned} \tag{15}$$

where $\mathbf{u}_2 = \big(u_2(0), u_2(1), \ldots\big) := c_2\big(\lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_3}\big)^{-1}$. We emphasize that the matrices $\mathbf{Q_3}$ and $\big(\lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_3}\big)$ given in (3) and (4) have a similar structure. Therefore, by analogy with the derivation of (8) in Lemma 4 we deduce that

$$\sum_{j_2 \geq i_2} z_2^{j_2} u_2(j_2) = z_2^{i_2} c_1 \big(\lambda_2 \hat{D}_2(z_2)\mathbf{I} + \lambda_3 \hat{D}_3(z_3)\mathbf{I} + \mathbf{A_2}\big)^{-1}. \tag{16}$$

Inserting (16) into (15) readily gives the desired result. □

We are now ready to report our main result for the autonomous-server discipline in the case $M = 3$.

Theorem 1. *The generating function of the joint queue-length of Q_1, Q_2 and Q_3 at the end of the server visit to Q_1 is given by*

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}^e_1}\big] = p(\mathbf{z})\mathbb{E}\big[r_1(z_2, z_3)^{N^b_{11}} z_2^{N^b_{21}} z_3^{N^b_{31}}\big] + q(\mathbf{z})\mathbb{E}\big[z_1^{N^b_{11}} z_2^{N^b_{21}} z_3^{N^b_{31}}\big], \tag{17}$$

where $\mathbf{z} := (z_1, z_2, z_3)$,

$$p(\mathbf{z}) = \frac{\alpha_1}{s_1(r_1(z_2, z_3), z_2, z_3)} \times \frac{(z_1 - 1)\tilde{B}_1(s_1(z_1, z_2, z_3))}{z_1 - \tilde{B}_1(s_1(z_1, z_2, z_3))}, \tag{18}$$

$$q(\mathbf{z}) = \frac{\alpha_1}{s_1(z_1, z_2, z_3)} \times \frac{z_1\big(1 - \tilde{B}_1(s_1(z_1, z_2, z_3))\big)}{z_1 - \tilde{B}_1(s_1(z_1, z_2, z_3))}, \tag{19}$$

$s_1(z_1, z_2, z_3) = \alpha_1 + \sum_{i=1}^3 \lambda_i(1 - \hat{D}_i(z_i))$, *and where* $r_1(z_2, z_3)$ *is the root with smallest absolute value of: (solving for* z_1*)*

$$z_1 = \tilde{B}_1\big(s_1(z_1, z_2, z_3)\big).$$

Proof. Multiplying (14) by $z_1^{j_1}$ and summing over all values of j_1 gives that

$$\begin{aligned}
\mathbb{E}\big[\,\mathbf{z}^{\mathbf{N}_1^e}\,\big|\mathbf{N}_1^b\big] &= \mathbb{E}\Big[\,z_1^{N_{11}^e} z_2^{N_{21}^e} z_3^{N_{31}^e}\,\Big|\mathbf{N}_1^b\Big] \\
&= -\alpha_1 z_2^{i_2} z_3^{i_3} c_1\big(\lambda_2\hat{D}_2(z_2)\mathbf{I}+\lambda_3\hat{D}_3(z_3)\mathbf{I}+\mathbf{A_2}\big)^{-1} \\
&\quad \times\big(\sum_{j_1\geq 0} z_1^{j_1} e_{j_1}\otimes e\big)^T \\
&= -\alpha_1 z_2^{i_2} z_3^{i_3}\Big(\sum_{j_1\geq 0} z_1^{j_1} u_1(j_1)\Big)e^T, \qquad (20)
\end{aligned}$$

where $\mathbf{u}_1 = \big(u_1(0), u_1(1), \ldots\big) := c_1\big(\lambda_2\hat{D}_2(z_2)\mathbf{I}+\lambda_3\hat{D}_3(z_3)\mathbf{I}+\mathbf{A_2}\big)^{-1}$. Let us now derive $\sum_{j_1\geq 0} z_1^{j_1} u_1(j_1)$. Note that $\mathbf{A}_2 = \mathbf{A}_1 - (\lambda_2+\lambda_3+\alpha_1)\mathbf{I}$ and $\mathbf{u}_1(\lambda_2\hat{D}_2(z_2)\mathbf{I}+\lambda_3\hat{D}_3(z_3)\mathbf{I}+\mathbf{A_2}\big) = c_1$. Inserting $\mathbf{A}_1$ given in (5) into the latter equation gives that

$$\begin{aligned}
-\theta u_1(0) + u_1(1)T_1^0 &= 0, \qquad (21)\\
\lambda_1 D_1(n)u_1(0)\pi_1 + \lambda_1\sum_{l=1}^{n-1} D_1(n-l)u_1(l)\mathbf{I} & \\
+u_1(n)(\mathbf{T}_1-\theta\mathbf{I}) + u_2(n+1)T_1^0\pi_1 &= \mathbf{1}_{\{n=i_1\}}\pi_1, \quad n\geq 1, \qquad (22)
\end{aligned}$$

where $\theta := \alpha_1+\lambda_1+\lambda_2(1-\hat{D}_2(z_2))+\lambda_3(1-\hat{D}_3(z_3))$. By multiplying (21) by π_1 and adding it to the sum over n of (22) multiplied by z_1^n, we find that

$$\sum_{n\geq 1} u_1(z_1)z_1^n\Big[\mathbf{T}_1 - \big(\theta-\lambda_1\hat{D}_1(z_1)\big)\mathbf{I} + \frac{1}{z_1}T_1^0\pi_1\Big] = \big[z_1^{i_1} + u_1(0)\big(\theta-\lambda_1\hat{D}_1(z_1)\big)\big]\pi_1. \qquad (23)$$

Let $\mathbf{R} := \big[\mathbf{T}_1 - \big(\theta-\lambda_1\hat{D}_1(z_1)\big)\mathbf{I} + \frac{1}{z_1}T_1^0\pi_1\Big]$. Then,

$$\sum_{n\geq 1} u_1(z_1)z_1^n = \big[z_1^{i_1} + u_1(0)\big(\theta-\lambda_1\hat{D}_1(z_1)\big)\big]\pi_1\mathbf{R}^{-1}. \qquad (24)$$

Inserting (24) into (20) we find that

$$\mathbb{E}\Big[\,z_1^{N_1^e} z_2^{N_2^e} z_3^{N_3^e}\,\Big|\mathbf{N}_1^b\Big] = -\alpha_1 z_2^{i_2} z_3^{i_3}\big(u_1(0) + \big[z_1^{i_1} + u_1(0)\big(\theta-\lambda_1\hat{D}_1(z_1)\big)\big]\pi_1\mathbf{R}^{-1}e^T\big), \qquad (25)$$

Now, we shall compute $\pi_1\mathbf{R}^{-1}e$. For the ease of the notation, let us denote $\mathbf{R}_1 := \mathbf{T}_1 - \big(\theta-\lambda_1\hat{D}_1(z_1)\big)\mathbf{I}$. By the Sherman-Morrison formula, see [3, Fact 2.14.2, p. 67], we have that

$$\begin{aligned}
\pi_1\mathbf{R}^{-1}e^T &= \pi_1\Big[\mathbf{R}_1^{-1} - \frac{1}{z_1}\big(1-\frac{1}{z_1}\tilde{B}_1(\theta-\lambda_1\hat{D}_1(z_1))\big)^{-1}\mathbf{R}_1^{-1}T_1^0\pi_1\mathbf{R}_1^{-1}\Big]e^T \\
&= \pi_1\mathbf{R}_1^{-1}e^T\Big[1+\frac{\frac{1}{z_1}\tilde{B}_1(\theta-\lambda_1\hat{D}_1(z_1))}{1-\frac{1}{z_1}\tilde{B}_1(\theta-\lambda_1\hat{D}_1(z_1))}\Big] \\
&= -\frac{1-\tilde{B}_1(\theta-\lambda_1\hat{D}_1(z_1))}{\theta-\lambda_1\hat{D}_1(z_1)} \times \frac{z_1}{z_1-\tilde{B}_1(\theta-\lambda_1\hat{D}_1(z_1))}, \qquad (26)
\end{aligned}$$

where the second equality follows from (1) and the last equality from Lemma 1. Inserting (26) into (25) yields that

$$\mathbb{E}\big[\, z_1^{N_1^e} z_2^{N_2^e} z_3^{N_3^e} \,\big|\, \mathbf{N}_1^b\big] = \frac{\alpha_1 z_1 z_2^{i_2} z_3^{i_3}[1-\tilde{B}_1(s_1(z_1,z_2,z_3))][z_1^{i_1}+u_1(0)s_1(z_1,z_2,z_3)]}{s_1(z_1,z_2,z_3)[z_1-\tilde{B}_1(s_1(z_1,z_2,z_3))]} - \alpha_1 z_2^{i_2} z_3^{i_3} u_1(0), \tag{27}$$

where $s_1(z_1,z_2,z_3)=\theta-\lambda_1\hat{D}_1(z_1)$. We shall show that for $|z_1|\le 1$ the denominator of (27) is not equal to zero except at one point. First, note that the real part of $\theta-\lambda_1\hat{D}_1(z_1)$ is strictly positive for $\alpha_1>0$, $|z_i|\le 1$, $i=1,2,3$. Moreover, by Rouché's theorem it is readily seen that $z_1-\tilde{B}_1(\theta-\lambda_1\hat{D}_1(z_1))=0$ has a unique root, $r_1(z_2,z_3)$, inside the unit the disk. Since the l.h.s. in (27) is a p.g.f. , it is analytical for $|z_1|\le 1$ we deduce that $r_1(z_2,z_3)$ is a removable singularity in (27), which gives

$$u_1(0) = -\frac{r_1(z_2,z_3)^{i_1}}{\theta-\lambda_1\hat{D}_1(r_1(z_2,z_3))}. \tag{28}$$

Inserting $u_1(0)$ into (27) and removing the condition on $\mathbf{N}_1^b$ readily gives $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_1^e}\big]$ in Theorem 1. ☐

General case. By analogy with the case of $M=3$, we order the transient states of $\mathbf{AMC}_A$ first according to n_M, then n_{M-1}, …, n_1, and finally according to ph_1. During a server visit to Q_1, the number of jobs at Q_j, $j=2,\dots,M$, may only increase. Therefore, similarly to the case of $M=3$, the $\mathbf{AMC}_A$ the generator matrix of the transition rates between the transient states of $\mathbf{AMC}_A$ for the general case, denoted by $\mathbf{Q}_M$, is an upper-triangular block matrix with diagonal blocks equal to $\mathbf{A}_M$, and i-th upper-diagonal blocks equal to $\lambda_M D_M(i)\mathbf{I}$. Moreover, $\mathbf{A}_M$ in turn is an upper-triangular block matrix with diagonal blocks equal to $\mathbf{A}_{M-1}$, and i-th upper-diagonal blocks equal to $\lambda_{M-1}D_{M-1}(i)\mathbf{I}$. We emphasize that $\mathbf{A}_j$, $j=M,\dots,3$, all verify the previous property. Finally, the matrix $\mathbf{A}_2=\mathbf{A}_1-(\lambda_2+\ldots+\lambda_M+\alpha_1)\mathbf{I}$, where $\mathbf{A}_1$ is the generator matrix of an M^X/PH/1 queue, with Poisson batch arrivals of inter-arrival rate λ_1 and batch size distribution function $D_1(\cdot)$.

By analogy with the $M=3$ case, we find that the probability of $\mathbf{N}_i^e=(j_1,\dots,j_M)$, given that $\mathbf{N}_1^b=(i_1,\dots,i_M)$, reads

$$\mathbb{P}\big(\mathbf{N}_1^e=(j_1,\dots,j_M)\mid \mathbf{N}_1^b\big) = -\alpha_1 c_M(\mathbf{Q}_M)^{-1}d_M, \tag{29}$$

where

$$c_M := e_{i_M}\otimes\ldots\otimes e_{i_1}\otimes\pi_1, \quad d_M := e_{j_M}\otimes\ldots\otimes e_{j_1}\otimes e.$$

Lemma 6. *The conditional generating function of the joint queue-length of $Q_2,\dots,Q_M$ at the end of the server visit to Q_1 is given by*

$$\mathbb{E}\Big[\prod_{i=2}^{M} z_i^{N_{i1}^e}\mathbf{1}_{\{N_{11}^e=j_1\}}\,\Big|\,\mathbf{N}_1^b\Big] = -\alpha_1\prod_{n=2}^{M} z_n^{i_n} c_1\Big(\sum_{i=2}^{M}\lambda_i\hat{D}_i(z_i)\mathbf{I}+\mathbf{A_2}\Big)^{-1}d_1^T.$$

Proof. Similar to the proof of Lemma 5. ☐

We are now ready to report our main result for the general case.

Theorem 2 (Autonomous-server discipline)**.** *The generating function of the joint queue-length of Q_1, …, Q_M at the end of the server visit to Q_1 is given by*

$$\gamma_1^A(\mathbf{z}) \quad = \quad p_1^A(\mathbf{z})\beta_1^A(\mathbf{z}_1^*) + q_1^A(\mathbf{z})\beta_1^A(\mathbf{z}_1), \tag{30}$$

where $\mathbf{z} = (z_1, \dots, z_M)$, $\mathbf{z}_1^* = \big(r_1(z_2, \dots, z_M), z_2, \dots, z_M\big)$,

$$p_1^A(\mathbf{z}) = \frac{\alpha_1}{s_1(\mathbf{z}_1^*)} \times \frac{(z_1 - 1)\tilde{B}_1(s_1(\mathbf{z}))}{z_1 - \tilde{B}_1(s_1(\mathbf{z}))}, \quad q_1^A(\mathbf{z}) = \frac{\alpha_1}{s_1(\mathbf{z})} \times \frac{z_1\big(1 - \tilde{B}_1(s_1(\mathbf{z}))\big)}{z_1 - \tilde{B}_1(s_1(\mathbf{z}))},$$

$s_1(\mathbf{z}) = \alpha_i + \sum_{i=1}^M \lambda_i(1 - \hat{D}_i(z_i))$, *and where* $r_1(z_2, \dots, z_M)$ *is the root with smallest absolute value of: (solving for* z_1*)*

$$z_1 = \tilde{B}_1\big(s_1(\mathbf{z})\big).$$

Proof. By analogy with the proof of Theorem 1. □

Eq. (30) relates $\gamma_1^A(\mathbf{z})$, the p.g.f. of the joint queue-length at the beginning of a server visit to Q_1, to $\beta_1^A(\mathbf{z}_1)$, the p.g.f. of the joint queue-length at the end of a server visit to Q_1. From Theorem 2, we deduce that for a server visit to Q_i, $i = 1, \dots, M$,

$$\gamma_i^A(\mathbf{z}) \quad = \quad p_i^A(\mathbf{z})\beta_i^A(\mathbf{z}_i^*) + q_i^A(\mathbf{z})\beta_i^A(\mathbf{z}_i), \tag{31}$$

where $\mathbf{z}_i^* = (z_1, \dots, z_{i-1}, r_i(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_M), z_{i+1}, \dots, z_M)$,

$$p_i^A(\mathbf{z}) = \frac{\alpha_i}{s_i(\mathbf{z}_i^*)} \times \frac{(z_i - 1)\tilde{B}_i(s_i(\mathbf{z}))}{z_i - \tilde{B}_i(s_i(\mathbf{z}))}, \quad q_i^A(\mathbf{z}) = \frac{\alpha_i}{s_i(\mathbf{z})} \times \frac{z_i\big(1 - \tilde{B}_i(s_i(\mathbf{z}))\big)}{z_i - \tilde{B}_i(s_i(\mathbf{z}))},$$

where $s_i(\mathbf{z}) = \alpha_i + \sum_{i=1}^M \lambda_i(1 - \hat{D}_i(z_i))$, and where $r_i(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_M)$ is the root with smallest absolute value of:

$$z_i = \tilde{B}_i\big(s_i(\mathbf{z})\big).$$

Finally, introducing the switch-over times from Q_{i-1} to Q_i, thus by using that $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^b}\big] = \mathbb{E}\big[\mathbf{z}^{\mathbf{N}_{i-1}^e}\big]C^{i-1}(\mathbf{z})$, where $C^{i-1}(\mathbf{z}) = C^{i-1}\Big(\sum_{i=1}^M \lambda_i\big(1 - \hat{D}_i(z_i)\big)\Big)$ is the p.g.f. of the number of Poisson batch arrivals during C^{i-1}, we obtain

$$\gamma_i^A(\mathbf{z}) \quad = \quad p_i^A(\mathbf{z})\gamma_{i-1}^A(\mathbf{z}_i^*)C^{i-1}(\mathbf{z}_i^*) + q_i^A(\mathbf{z})\gamma_{i-1}^A(\mathbf{z})C^{i-1}(\mathbf{z}). \tag{32}$$

Remark 1. *In the particular case where* $\hat{D}_i(z_i) = z_i$*, i.e., the arriving batches are all of size one, Eq. (31) agrees with* [7, Theorem 5.3].

4. Time-Limited discipline

In this section, we will relate the joint queue-length probabilities at the beginning and end of a server visit to a queue for the time-limited discipline. Under this discipline, the server departs from Q_i when it becomes empty or when a timer of exponentially duration with rate α_i has expired, whichever occurs first. Moreover, if the server arrives to an empty queue, he leaves the queue immediately and jumps to the next queue in the schedule. For this reason, we should differentiate here between the two events where the server joins an empty and non-empty queue.

We will follow the same approach as in Section 3. Thus, we first assume that there are $\mathbf{N}_1^b := (i_1, ..., i_M)$ jobs in $(Q_1, \dots, Q_M)$, with $i_1 \geq 1$, at the beginning time of a server visit to Q_1 and second there are $\mathbf{N}_1^e := (\mathbf{N}_{11}^e, ..., \mathbf{N}_{1M}^e) = (j_1, ..., j_M)$ jobs in $(Q_1, \dots, Q_M)$ at the end time of a server visit to Q_1. Note that if Q_1 is empty at the beginning of a server visit, i.e., $i_1 = 0$, then $\mathbb{P}(\mathbf{N}_1^e = \mathbf{N}_1^b) = 1$. We shall exclude the latter obvious case from the analysis in the following. However, we shall include it when the result is unconditioned on $\mathbf{N}_1^b$.

Let $\mathbf{N}(t) := (PH_1(t), N_1(t), \ldots, N_M(t))$ denote the $(M+1)$-dimensional, continuous-time Markov chain with discrete state-space $\xi_T = \{1, \ldots, h_1\} \times \{0, 1, \ldots\}^M \cup \{a\}$, where $N_j(t)$ represents the number of jobs in Q_j at time t and at which Q_1 is being served. State $\{a\}$ is absorbing. We refer to this absorbing Markov chain by $\mathbf{AMC}_T$. The absorption of $\mathbf{AMC}_T$ occurs when the server leaves Q_1 which happens with rate α_1 from all transient states. The transient states of the form $(ph_1, 1, n_2, \ldots, n_M)$ have an additional transition rate to $\{a\}$ that is equal to the (ph_1)-entry of T_1^0 which represents the departure of the last job at Q_1 from the service phase ph_1.

We shall now derive the joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to timer expiration and later the joint conditional p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to Q_1 empty. We set $\mathbf{N}(0) = (PH_1(0), \mathbf{N}_1^b)$, where $PH_1(0)$ is distributed according to π_1, i.e., preemptive repeat discipline. We order the transient states lexicographically first according to n_M, then to n_{M-1},…, n_1, and finally according to ph_1. Similarly to the autonomous-server discipline, during a server visit to Q_1, the number of jobs at Q_j, $j = 2, \ldots, M$, may only increase. It then follows that the transient generator of $\mathbf{AMC}_T$ has the same structure as the transient generator of $\mathbf{AMC}_A$, i.e. it is an upper-triangular Toeplitz matrix of upper-triangular Toeplitz diagonal blocks. Therefore, by the same arguments as for the autonomous-server, we find that the joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to timer expiration, denoted by $\{\text{timer}\}$, given $\mathbf{N}_1(0)$, reads

$$\mathbb{E}\Big[\, \mathbf{z}^{\mathbf{N}_1^e} \mathbf{1}_{\{\text{timer}\}} \,\Big|\, \mathbf{N}_1^b \Big] = -\alpha_1 \prod_{n=2}^{M} z_n^{i_n} c_1 \Big(\sum_{i=2}^{M} \lambda_i \hat{D}_i(z_i)\mathbf{I} + \mathbf{B_2} \Big)^{-1} g_1(z_1)^T, \tag{33}$$

where $\mathbf{B}_2 := \mathbf{B}_1 - (\lambda_2 + \ldots + \lambda_M + \alpha_1)\mathbf{I}$, $\mathbf{B}_1$ is the generator matrix of an $\mathrm{M}^X/\mathrm{PH}/1$ queue restricted to the states with the number of jobs strictly positive, i.e., $\mathbf{B}_1$ is obtained by deleting the first row of blocks and column of the matrix $\mathbf{A}_1$ defined in (5), and where

$$g_1(z_1) \quad := \quad \sum_{j_1 \geq 1} z_1^{j_1} e_{j_1} \otimes e = (z_1 e, z_1^2 e, \ldots), \quad c_1 = e_{i_1} \otimes \pi_1.$$

Let $\mathbf{Q_T}(\mathbf{z}) = \sum_{i=2}^{M} \lambda_i \big(1 - \hat{D}_i(z_i)\big)\mathbf{I} + \mathbf{B_1}$.

Lemma 7. *The joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to timer expiration, given $\mathbf{N}_1^b = (i_1, \ldots, i_M)$, is given by*

$$\mathbb{E}\Big[\, \mathbf{z}^{\mathbf{N}_1^e} \mathbf{1}_{\{timer\}} \,\Big|\, \mathbf{N}_1^b \Big] = \alpha_1 z_1 \Big(\prod_{n=2}^{M} z_n^{i_n} \Big) \frac{[z_1^{i_1} - r_1(z_2, \ldots, z_M)^{i_1}][1 - \tilde{B}_1\big(s_1(\mathbf{z})\big)]}{s_1(\mathbf{z})[z_1 - \tilde{B}_1\big(s_1(\mathbf{z})\big)]}, \tag{34}$$

where $r_1 = \tilde{B}_1\big(s_1(r_1, z_2, \ldots, z_M)\big)$ and $s_1(\mathbf{z}) = \alpha_1 + \sum_{i=1}^{M} \lambda_i (1 - \hat{D}_i(z_i))$.

Proof. Equation (33) yields that

$$\mathbb{E}\Big[\mathbf{z}^{\mathbf{N}_1^e} \mathbf{1}_{\{\text{timer}\}} \mid \mathbf{N}_1^b\Big] \quad = \quad -\alpha_1 \prod_{n=2}^{M} z_n^{i_n} \Big(\sum_{j_1 \geq 1} z_1^{j_1} u_1(j_1) \Big) e^T, \tag{35}$$

where $\mathbf{u}_1 = \big(u_1(1), u_1(2), \dots\big) := c_1(\mathbf{Q_T}(\mathbf{z}))^{-1}$. Note that $\mathbf{u}_1\mathbf{Q_T}(\mathbf{z}) = c_1$. Inserting $\mathbf{Q_T}(\mathbf{z})$ into the latter equation gives that

$$\mathbf{1}_{\{n\geq 2\}}\lambda_1 \sum_{l=1}^{n-1} D_1(n-l)u_1(l)\mathbf{I} + u_1(n)(\mathbf{T}_1 - \theta\mathbf{I}) + u_2(n+1)T_1^0\pi_1 = \mathbf{1}_{\{n=i_1\}}\pi_1, \tag{36}$$

where $\theta = \alpha_1 + \lambda_1 + \sum_{i=2}^M \lambda_i(1-\hat{D}_i(z_i))$. Multiplying (36) by z_1^n and summing over n yields that

$$\sum_{n\geq 1} u_1(z_1)z_1^n = \big[z_1^{i_1} + u_1(1)T_1^0\big]\pi_1\mathbf{R}^{-1}. \tag{37}$$

Inserting (37) into (35) we find that

$$\begin{aligned}\mathbb{E}\Big[\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{\text{timer}\}} \mid \mathbf{N}_1^b\Big] &= -\alpha_1\Big(\prod_{n=2}^M z_n^{i_n}\Big)\big[z_1^{i_1} + u_1(1)T_1^0\big]\pi_1\mathbf{R}^{-1}e^T \\ &= \alpha_1 z_1\Big(\prod_{n=2}^M z_n^{i_n}\Big)\frac{[z_1^{i_1} + u_1(1)T_1^0][1-\tilde{B}_1\big(s_1(\mathbf{z})\big)]}{s_1(\mathbf{z})[z_1 - \tilde{B}_1\big(s_1(\mathbf{z})\big)]},\end{aligned} \tag{38}$$

where the second equality follows from (26) and $s_1(\mathbf{z}) = \theta - \lambda_1\hat{D}_1(z_1)$. Because the joint moment generating function $\mathbb{E}\Big[\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{\text{timer}\}} \mid \mathbf{N}_1^b\Big]$ in (38) has a singular point at $z_1 = r_1(z_2,\dots,z_M)$, $|r_1(z_2,\dots,z_M)| < 1$, it should be removable. Thus,

$$u_1(1)T_1^0 = -r_1(z_2,\dots,z_M)^{i_1}, \tag{39}$$

where $r_1(z_2,\dots,z_M) = \tilde{B}_1\big(s_1(r_1(z_2,\dots,z_M), z_2,\dots,z_M)\big)$. Inserting $u_1(1)T_1^0$ into (38) readily gives $\mathbb{E}\Big[\,\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{\text{timer}\}}\,\Big|\,\mathbf{N}_1^b\Big]$. □

Lemma 8. *The joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to empty Q_1, given $\mathbf{N}_1^b = (i_1,\dots,i_M)$, is given by*

$$\mathbb{E}\Big[\,\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{timer\}}\,\Big|\,\mathbf{N}_1^b\Big] = r_1(z_2,\dots,z_M)^{i_1}\prod_{n=2}^M z_n^{i_n}, \tag{40}$$

where $r_1(z_2,\dots,z_M) = \tilde{B}_1\big(s_1(r_1(z_2,\dots,z_M), z_2,\dots,z_M)\big)$ and $s_1(\mathbf{z}) = \alpha_1 + \sum_{i=1}^M \lambda_i(1-\hat{D}_i(z_i))$.

Proof. The joint moment of the p.g.f. of $\mathbf{N}_1^e$ and the event that the absorption is due to Q_1 being empty, is given by

$$\begin{aligned}\mathbb{E}\Big[\,\mathbf{z}^{\mathbf{N}_1^e}\mathbf{1}_{\{Q_1\text{ empty}\}}\,\Big|\,\mathbf{N}_1^b\Big] &= -\prod_{n=2}^M z_n^{i_n}c_1\mathbf{Q_T}(\mathbf{z})^{-1}e_1^T \otimes T_1^0 \\ &= -\prod_{n=2}^M z_n^{i_n}u_1(1)T_1^0 \\ &= r_1(z_2,\dots,z_M)^{i_1}\prod_{n=2}^M z_n^{i_n},\end{aligned}$$

where $\mathbf{u}_1 = c_1(\mathbf{Q_T}(\mathbf{z}))^{-1}$ and the last equality follows from (39). □

Combining Lemmas 7 and 8 we obtain our main theorem for the time-limited discipline.

Theorem 3 (Time-limited discipline)**.** *The generating function of the joint queue-length of Q_1, ..., Q_M at the end of the server visit to Q_1 is given by*

$$\gamma_1^T(\mathbf{z}) \;=\; p_1^T(\mathbf{z})\beta_1^T(\mathbf{z}_1^*) + q_1^T(\mathbf{z})\beta_1^T(\mathbf{z}),$$

where $\mathbf{z} = (z_1, \ldots, z_M)$, $\mathbf{z}_1^* = \big(r_1(z_2, \ldots, z_M), z_2, \ldots, z_M\big)$,

$$p_1^T(\mathbf{z}) = 1 - \frac{\alpha_1}{s_1(\mathbf{z})} \times \frac{z_1\big(1 - \tilde{B}_1(s_1(\mathbf{z}))\big)}{z_1 - \tilde{B}_1(s_1(\mathbf{z}))}, \quad q_1^T(\mathbf{z}) = \frac{\alpha_1}{s_1(\mathbf{z})} \times \frac{z_1\big(1 - \tilde{B}_1(s_1(\mathbf{z}))\big)}{z_1 - \tilde{B}_1(s_1(\mathbf{z}))},$$

where $s_1(\mathbf{z}) = \alpha_i + \sum_{i=1}^M \lambda_i(1 - \hat{D}_i(z_i))$ *and* $r_1(z_2, \ldots, z_M)$ *is the root with smallest absolute value of: (solving according to* z_1*)*

$$z_1 = \tilde{B}_1\big(s_1(\mathbf{z})\big).$$

We deduce that for a server visit to Q_i, $i = 1, \ldots, M$,

$$\gamma_i^T(\mathbf{z}) \;=\; p_i^T(\mathbf{z})\beta_i^T(\mathbf{z}_i^*) + q_i^T(\mathbf{z})\beta_1^T(\mathbf{z}), \tag{41}$$

where $\mathbf{z}_i^* = (z_1, \ldots, z_{i-1}, r_i(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M), z_{i+1}, \ldots, z_M)$,

$$p_i^T(\mathbf{z}) = 1 - \frac{\alpha_1}{s_1(\mathbf{z})} \times \frac{z_1\big(1 - \tilde{B}_1(s_1(\mathbf{z}))\big)}{z_1 - \tilde{B}_1(s_1(\mathbf{z}))}, \quad q_i^T(\mathbf{z}) = \frac{\alpha_1}{s_1(\mathbf{z})} \times \frac{z_1\big(1 - \tilde{B}_1(s_1(\mathbf{z}))\big)}{z_1 - \tilde{B}_1(s_1(\mathbf{z}))},$$

where $s_i(\mathbf{z}) = \alpha_i + \sum_{i=1}^M \lambda_i(1 - \hat{D}_i(z_i))$, and where $r_i(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M)$ is the root with smallest absolute value of:

$$z_i = \tilde{B}_i\big(s_i(\mathbf{z})\big).$$

Finally, introducing the switch-over times from Q_{i-1} to Q_i, thus by using that $\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^b}\big] = \mathbb{E}\big[\mathbf{z}^{\mathbf{N}_{i-1}^e}\big]C^{i-1}(\mathbf{z})$, where $C^{i-1}(\mathbf{z})$ is the p.g.f. of the number of Poisson batch arrivals during C^{i-1}, we obtain

$$\gamma_i^T(\mathbf{z}) \;=\; p_i^T(\mathbf{z})\gamma_{i-1}^T(\mathbf{z}_i^*)C^{i-1}(\mathbf{z}_i^*) + q_i^T(\mathbf{z})\gamma_{i-1}^T(\mathbf{z})C^{i-1}(\mathbf{z}). \tag{42}$$

Remark 2. *In the particular case where* $\hat{D}_i(z_i) = z_i$*, i.e. the arriving batches are all of size one, Eq. (41) agrees with* [7, Theorem 5.10].

Remark 3. Exhaustive discipline. *Taking the limit of (41) for* $\alpha_i \to 0$ *the time-limited discipline is equivalent to the exhaustive discipline. We find that*

$$\mathbb{E}\big[\mathbf{z}^{\mathbf{N}_i^e}\big] = \mathbb{E}\big[(\mathbf{z}_i^*)^{\mathbf{N}_i^b}\big], \tag{43}$$

where $\mathbf{z}_i^* := (z_1, \ldots, z_{i-1}, y_i, z_{i+1}, \ldots, z_M)$ *and* y_i *is the root of*

$$z_i = \tilde{B}_i\Big(\sum_{i=1}^M \lambda_i(1 - \hat{D}_i(z_i))\Big). \tag{44}$$

Eq. (43) is equivalent to the well-known relation of exhaustive discipline in (see, e.g., [9, Eq. (24)]*).*

5. Iterative scheme

In this section, we will explain how to obtain the joint queue-length distribution using an iterative scheme. First, we obtain $\gamma_i(\mathbf{z})$ as function $\gamma_{i-1}(\mathbf{z})$, where $\mathbf{z} = (z_1, \ldots, z_M)$.

Note that $\gamma_i(\mathbf{z})$ is a function of $\gamma_{i-1}(\mathbf{z})$ and $\gamma_{i-1}(\mathbf{z}_i^*)$ where $\mathbf{z}_i^* = (z_1, \ldots, z_{i-1}, a, z_{i+1} \ldots, z_M)$ with $|z_i| = 1$, $i = 1, \ldots, M$ and $|a| \leq 1$. Moreover, we note that a is function of z_l for all $l = 1, \ldots, M$ and $l \neq i$. Since $\gamma_{i-1}(\mathbf{z})$ is a p.g.f. it should be analytic in z_i for all $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M$. Hence, we can write

$$\gamma_{i-1}(\mathbf{z}) = \sum_{n=0}^{\infty} g_{in}(z_1, \ldots, z_{i-1}, z_{i+1} \ldots, z_M) z_i^n, \quad |z_i| \leq 1,$$

where $g_{in}(.)$ is again an analytic function. From complex function theory, it is well known that (see, e.g., [19])

$$\gamma_{i-1}(\mathbf{z}_i^*) = \frac{1}{2\pi \mathbf{i}} \oint_C \frac{\gamma_{i-1}(\mathbf{z})}{z_i - a} dz_i, \quad |a| \leq 1,$$

where C is the unit circle and $\mathbf{i}^2 = -1$. In addition, we have that

$$g_{in}(z_1, \ldots, z_{i-1}, z_{i+1} \ldots, z_M) = \frac{1}{2\pi \mathbf{i}} \oint_C \frac{\gamma_{i-1}(\mathbf{z})}{z_i^{n+1}} dz_i,$$

where $n = 0, 1, \ldots$. These formulas show that we only need to know the p.g.f. $\gamma_{i-1}(\mathbf{z})$ for all $\mathbf{z}$ with $|z_i| = 1$, to be able to compute $\gamma_i(\mathbf{z})$.

When there is an incurred switch-over time from queue $i-1$ to i the p.g.f. of the joint queue-length at the end of the n-th server visit to Q_i, denoted by $\gamma_i^n(\mathbf{z})$, can be computed as function of $\gamma_{i-1}^n(\mathbf{z})$, see Eq. (32) and (42). The main step is to iterate over all queues in order to express $\gamma_i^{n+1}(\mathbf{z})$ as function of $\gamma_i^n(\mathbf{z})$. Assuming that the system is in steady-state these two latter quantities should be equal. Thus, starting with an empty system at the first service visit to Q_i and repeating the latter main step we can compute $\gamma_i^2(\mathbf{z})$, $\gamma_i^3(\mathbf{z})$, and so on. This iteration is stopped when $\gamma_i^n(\mathbf{z})$ converges.

6. Discussion and Conclusion

In this paper, we have developed a general framework to analyze polling systems with Poisson batch arrivals and phase-type service times for the autonomous-server and the time-limited service discipline. The framework is based on the key idea of relating directly the joint queue-length distribution at the beginning and the end of a server visit. In order to do so, we used the theory of absorbing Markov chains. We have illustrated our framework for the autonomous-server and the time-limited service discipline. The analysis presented in this paper is restricted to the case of a single job service at a time. We emphasize that the analysis can be extended to the more general batch service disciplines, see [6, Chap. III.2]. For instance, Lemma 6 holds in this case, however, the matrix $\mathbf{A_2}$ becomes a full block matrix.

In this paper we have showed that our framework is applicable to disciplines that do not satisfy the branching property that are, in general, considered to be hard to analyze. Our framework is also applicable to branching type polling systems such as the exhaustive discipline. Moreover, we claim that with an extra effort one can analyze the gated discipline for which there already exist results in the literature.

Acknowledgements

In the Netherlands, the 3 universities of technology have formed the 3TU.Federation. This article is the result of joint research in the 3TU.Centre of Competence NIRICT (Netherlands Institute for Research on ICT). The authors would thank De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) for their financial support.

References

[1] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. A tandem queueing model for delay analysis in disconnected ad hoc networks. In *Proc. of ASMTA, LCNS 5055*, pages 189–205, Nicosia, Cyprus, June 2008.
[2] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. Time-limited and k-limited polling systems: A matrix analytic solution. In *Proc. of SMCTools*, Athens, Greece, Oct. 2009.
[3] D. S. Bernstein. *Matrix Mathematics.* Princeton University Press, 2005.
[4] J. Blanc. An algorithmic solution of polling models with limited service disciplines. 40(7):1152–1155, July 1992.
[5] J. Blanc. The power-series algorithm for polling systems with time limits. *Probability in the Engineering and Informational Sciences*, 12:221–237, 1998.
[6] J. W. Cohen. *The single server queue.* North-Holland, 1982.
[7] R. de Haan. *Queueing models for mobile ad hoc networks.* PhD thesis, Enschede, June 2009. http://doc.utwente.nl/61385/.
[8] R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. A polling model with an autonomous server. *Research Memorandum 1845, University of Twente*, 2007.
[9] M. Eisenberg. Queues with periodic service and changeover times. *Operation Research*, 20(2):440–451, 1972.
[10] D. P. Gaver, P. A. Jacobs, and G. Latouche. Finite birth-and-death models in randomly changing environments. *Adv. Appl. Probab.*, 16:715–731, 1984.
[11] C. Grinstead and J. Snell. *Introduction to Probability.* American Mathematical Society, 1997.
[12] F. Guillemin and A. Simonian. Transient characteristics of an M/M/1/infinity system. *Advances in Appl. Prob.*, 27:862–888, 1995.
[13] K. Leung. Cyclic-service systems with probabilistically-limited service. *IEEE Journal on Selected Areas in Communications*, 9(2):185–193, 1991.
[14] K. Leung. Cyclic-service systems with non-preemptive time-limited service. *IEEE Transactions on Communications*, 42(8):2521–2524, 1994.
[15] H. Levy and M. Sidi. Polling systems: Applications, modeling, and optimization. *TOC*, 38(10), Oct. 1990.
[16] M. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Johns Hopkins University Press, 1981.
[17] J. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(10):409–429, 1993.
[18] H. Takagi. Analysis and application of polling models. In *Performance Evaluation: Origins and Directions, LNCS 1769*, pages 423–442, Berlin, Germany, 2000. Springer-Verlag.
[19] E. Titchmarsh. *The Theory of Functions.* Oxford Science Publications, 1976.
[20] M. van Vuuren and E. Winands. Iterative approximation of k-limited polling systems. *Queueing Systems: Theory and Applications*, Vol. 55(3):161 – 178, 2007.
[21] U. Yechiali and I. Eliazar. Polling under the randomly-timed gated regime. *Stochastic models*, 14(1):79–93, 1998.