Decomposition results for general polling systems and their applications

Dimitris Bertsimas [a] and Georgia Mourtzinou [b]

[a] *Massachusetts Institute of Technology, Cambridge, MA 02142, USA*
E-mail: dbertsim@mit.edu
[b] *Dynamic Ideas, LLC., Cambridge, MA 02142, USA*
E-mail: gina@aris.mit.edu

Received 23 July 1997; revised 10 December 1998

In this paper we derive decomposition results for the number of customers in polling systems under arbitrary (dynamic) polling order and service policies. Furthermore, we obtain sharper decomposition results for both the number of customers in the system and the waiting times under static polling policies. Our analysis, which is based on distributional laws, relaxes the Poisson assumption that characterizes the polling systems literature. In particular, we obtain exact decomposition results for systems with either Mixed Generalized Erlang (MGE) arrival processes, or asymptotically exact decomposition results for systems with general renewal arrival processes under heavy traffic conditions. The derived decomposition results can be used to obtain the performance analysis of specific systems. As an example, we evaluate the performance of gated Markovian polling systems operating under heavy traffic conditions. We also provide numerical evidence that our heavy traffic analysis is very accurate even for moderate traffic.

Keywords: polling systems, switch-over times, decomposition, distributional laws, heavy traffic, mixed generalized Erlang arrivals, performance analysis

1. Introduction

Polling systems were introduced in the early 1970s as models of time-sharing computer systems (Takagi [24] indicates that polling goes back to the patrolling repairman problem in the 1950s). Currently, they are extensively used to model queueing systems where many job classes share a single server and a setup time is incurred whenever the server changes classes. Such systems represent a broad range of applications: production and manufacturing systems, traffic and transportation systems and, as we already mentioned, computer and communications systems.

In this paper we study polling systems with general renewal arrival processes for the different job classes, arbitrary polling order and service policies in the individual queues, and we address the following questions: Are there structural relationships that prevail in such a general setting? Is there a subclass of polling systems for which we have sharper, easier to use, results? If so, how can our structural results be used to yield the performance analysis of specific systems?

Traditionally, the literature in the area of polling systems deals with the performance analysis of specific models under Poisson arrivals. The earliest papers on the topic considered models of cyclic polling (see, for example, Konheim and Meister [20] for a discrete time model, and Cooper and Murray [10], Eisenberg [12] for continuous time models). Recently, more general polling systems have been proposed and analyzed in Kleinrock and Levy [19], Boxma and Weststrate [9], and in Baker and Rubin [2]. A thorough survey of polling systems may be found in Takagi [24] and more recent results appear in Takagi [25] and Levy and Sidi [22]. There are also a few papers dealing with general inputs for cyclic polling systems, for example, Tran-Gia [26] (for a discrete time model) and Bertsimas and Mourtzinou [4] (for a continuous time model).

This paper follows a more recent trend in the polling systems literature, establishing structural results for a variety of polling models with Poisson arrivals. Fuhrmann [13] and Fuhrmann and Cooper [14] first established decomposition results for cyclic polling systems. Based on those decomposition results, Boxma and Groenendijk [8] proved that the total amount of work in a cyclic polling system is composed of two independent components: one is the amount of work in the corresponding system without the switch-over times, the other is the amount of work at an arbitrary epoch during a switch-over period; those results are known as pseudo-conservation laws. Similar pseudo-conservation laws hold for a variety of polling systems, including systems with probabilistic polling order and systems with polling tables, as illustrated in a survey paper by Boxma [7]. Recently, Srinivasan et al. [23] related the waiting time distributions in a cyclic polling system with nonzero switch-over time to the corresponding distributions in a polling system with zero switch-overs via a decomposition argument and then derived explicit relationships for the waiting time moments. Their results were extended in Borst and Boxma [6].

The contributions of this work are as follows:

1. We derive a set of decomposition results for the number of customers in polling systems, which allow for arbitrary polling order and service policies. Furthermore, we identify sufficient conditions on the systems, namely static polling policies, that give rise to sharper decomposition results for both the number of customers in the queue and the waiting times.

2. Our decomposition results relate the number of customers and the waiting times in the *individual* queues of the polling system and the corresponding quantities in the same queues *in isolation*. Therefore, they can easily be used to obtain the performance analysis of each queue of the polling system, in contrast with pseudo-conservation laws that can only yield bounds for the individual queues. To emphasize this point we obtain, in the last part of the paper, the performance analysis of Markovian polling systems with general arrival and service distributions, under heavy traffic conditions. We also provide numerical evidence that our

heavy traffic analysis gives rise to very accurate results, compared with simulation, even for moderate traffic.

3. We relax the Poisson assumption that characterizes the polling systems literature. In particular we obtain decomposition results for systems with either Mixed Generalized Erlang (MGE) arrival processes, or general renewal arrival processes provided that the system is operating under heavy traffic conditions.

Regarding the methodological contribution of this work, we propose a new method for addressing polling systems based on distributional laws first obtained by Haji and Newell [15]. Our work in this paper further demonstrates the significance of these laws as already noted by Keilson and Servi [16,17] and Bertsimas and Mourtzinou [3,4].

The rest of the paper is structured as follows. In section 2, we present the model and we introduce the necessary notation. Next, in sections 3 and 4, we prove the decomposition results for systems with mixed generalized Erlang arrivals and for systems with renewal arrivals under heavy traffic, respectively. In section 5, we apply our analysis to obtain the performance of a Markovian polling system under heavy traffic and we compare our results to simulation experiments. Finally, we conclude with some remarks and future directions in section 6.

2. Model description and notation

We consider a general queueing system, in which a single server is servicing N classes of customers. Class i customers arrive at queue (node) i according to a renewal arrival process described by $N_{a_i}(t)$, the number of arrivals in the interval $(0, t]$, with mean interarrival time $1/\lambda_i$. We denote by $N^*_{a_i}(t)$ the number of arrivals in the interval $(0, t]$ from the equilibrium arrival process. The service time of class i customers, represented by X_i, follows a general distribution with mean $E[X_i]$. We impose the following assumptions:

A.1. All arriving customers enter the system one at a time, remain in the system until served (there is no blocking, balking or reneging) and leave also one at a time.

A.2. The customers within each class leave the system in the order of arrival (FIFO).

A.3. For each class, new arriving customers do not affect the time in the system of previous customers.

A.4. Different arrival processes are mutually independent.

The above assumptions are commonly used in the literature and allow for dependencies between the service times of different customer classes.

Regarding the polling policy, we start our analysis by allowing the server to poll the different nodes using an arbitrary policy and to encounter a random delay d_{ij} every time he switches from node i to node j. If all switch-over times are deterministic and equal to zero, we further assume (see, for example, [24]) that the server stops switching whenever the system becomes empty, and then instantaneously switches to the queue where the next customer arrives. The reason we impose this further assumption is that we do not want the server to switch infinitely many times between queues in zero time. We do not impose any conditions on the way the server is servicing the individual nodes other than assumption A.2.

Next, to obtain sharper decomposition results, we focus on a subclass of polling systems where:

A.5. The polling order is independent of the number of customers in the system.

A.6. The service policy is gated for classes in a set E_g and exhaustive for the rest of classes.

A.7. The switch-over times are nonzero.

We refer to polling policies satisfying assumption A.5 as *static polling policies*. The class of static policies contains the majority of polling policies studied in the literature, i.e., cyclic or probabilistic polling as well as fixed-order polling tables. Furthermore, assumption A.6 focuses on the two most commonly used service policies, exhaustive and gated. Let us describe those service policies: If queue i is served in a gated mode and there are N_i customers waiting in the queue when the server starts servicing this class, then the server processes all N_i customers in a FIFO order, and then polls queue j – according to the general policy – after encountering a random delay d_{ij}. Notice that the class i customers that arrive while the server is servicing the N_i customers have to wait for the next visit of the server to the ith queue, i.e., for a full cycle to be completed. On the other hand, if queue j is served in an exhaustive mode, then whenever the server polls this queue it continues servicing in a FIFO order until the queue empties. We introduce the following notation:

T_i^k: the time that the server spends servicing the ith class in the kth visit;

C_i^k: the $(k-1)$st cycle with respect to class i, i.e., the time interval from the $(k-1)$st entrance to queue i until the kth entrance to queue i;

Δ_i^k: the intervisit time with respect to class i, i.e., the time between the end of the $(k-1)$st visit and the beginning of the kth visit to class i;

C_{ij}^k: the time interval between the kth entrance to queue i and the previous entrance to queue j.

Notice that $C_{ii}^k = C_i^k$. Further, we let ρ_i be the utilization of station i, $\rho_i \triangleq \lambda_i E[X_i]$, and $\rho \triangleq \sum_{i=1}^N \rho_i$ be the total traffic intensity. We assume throughout this paper that stability conditions are fulfilled and that the system reaches steady-state (the ergodicity conditions depend on the service policies at the queues; obviously $\rho < 1$ is a *necessary* condition, but not always sufficient).

We further define the following quantities in steady-state:

Q_i: the number of customers waiting in queue i in steady-state;

L_i: the number of customers of class i in the system (queue i plus server) in steady-state;

W_i: the steady-state waiting time of a class i customer;

S_i: the steady-state system time, i.e., the waiting time plus the service time, of a class i customer;

$$C_i \triangleq \lim_{k\to\infty} C_i^k, \quad \Delta_i \triangleq \lim_{k\to\infty} \Delta_i^k, \quad C_{ij} \triangleq \lim_{k\to\infty} C_{ij}^k.$$

We also denote by $E[z^Y]$ the z-transform of the discrete random variable Y, and by $\phi_Z(s)$ the Laplace–Stieltjes transform of the random variable Z, i.e.,

$$\phi_Z(s) \triangleq \int_0^\infty \mathrm{e}^{-st}\, \mathrm{d}F_Z(t).$$

Finally, we define $F_Z(t) \triangleq Pr\{Z \leqslant t\}$ the distribution function of Z.

3. Decomposition results for polling systems with MGE arrivals

In this section we establish decomposition results for polling systems with MGE arrival processes. The MGE distribution, i.e., the Coxian distribution (see [11]) with real rates, is used very often in practice since it is the simplest class of distributions that is dense in the space of all distributions, i.e., it can approximate any renewal arrival process arbitrarily closely. The MGE arrival process with M_i stages, denoted by MGE$_{M_i}$, can be represented as an arrival timing channel (ATC) consisting of M_i consecutive exponential stages with rates $\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,M_i}$ and with probabilities $p_{i,1}, p_{i,2}, \ldots, p_{i,M_i}$ ($p_{i,M_i} = 1$) of entering the system after the completion of the 1st, 2nd, $\ldots, M_i$th stage. By introducing the following upper semi-diagonal matrix $A_{i,0}$ and dyadic matrix $A_{i,1}$:

$$A_{i,0} = \begin{bmatrix} \lambda_{i,1} & -(1-p_{i,1})\lambda_{i,1} & 0 & \ldots & 0 \\ 0 & \lambda_{i,2} & -(1-p_{i,2})\lambda_{i,2} & \ddots & \vdots \\ \vdots & \ddots & & \ddots & \vdots \\ \vdots & & & \lambda_{i,M_i-1} & -(1-p_{i,M_i-1})\lambda_{i,M_i-1} \\ 0 & \ldots & \ldots & 0 & \lambda_{i,M_i} \end{bmatrix},$$

$$A_{i,1} = \begin{bmatrix} -p_{i,1}\lambda_{i,1} & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ -p_{i,M_i}\lambda_{i,M_i} & 0 & \ldots & 0 \end{bmatrix},$$

we can express the interarrival pdf as

$$a_i(t) = -\mathrm{trace}\bigl(\mathrm{e}^{-A_{i,0}t} A_{i,1}\bigr) \quad \text{and} \quad \lambda_i \triangleq -\frac{1}{a_i(0)}.$$

We further define $\boldsymbol{e}_1 \overset{\Delta}{=} (1,0,\dots,0)$, $\mathbf{1} \overset{\Delta}{=} (1,\dots,1,\dots,1)$, R_i to be the ATC stage and

$$\boldsymbol{P}_{i,n} = \bigl[P\{L_i = n,\ R_i = i\}\bigr]_{i=1}^{i=M_i} \quad \text{and} \quad \boldsymbol{P}_{L_i}(z) = \sum_{n=0}^{\infty} z^n \boldsymbol{P}_{i,n}.$$

For systems with MGE arrival processes Bertsimas and Nakazato in [5] proved the following vector distributional law.

Theorem 1 (Bertsimas and Nakazato [5]). Under assumptions A.1–A.3 and for mixed generalized Erlang interarrival times characterized by the matrices $A_{i,0}$, $A_{i,1}$,

$$\boldsymbol{P}_{L_i}(z) = \lambda_i(1-z)\,\boldsymbol{e}_1\Phi_{S_i}(A_{i,0}+zA_{i,1})(A_{i,0}+zA_{i,1})^{-1}, \tag{1}$$

where for any matrix D we symbolically define: $\Phi_S(D) \overset{\Delta}{=} \int_0^\infty \mathrm{e}^{-Dt}\,\mathrm{d}F_S(t)$.

If we define the system to be only queue i of the polling system, equation (1) becomes

$$\boldsymbol{P}_{Q_i}(z) = \lambda_i(1-z)\,\boldsymbol{e}_1\Phi_{W_i}(A_{i,0}+zA_{i,1})(A_{i,0}+zA_{i,1})^{-1}. \tag{2}$$

We next prove a decomposition theorem for polling systems with MGE arrivals that allows for an arbitrary polling policy that may depend on the number of customers in each queue. Let us denote by B_i the event that at a random observation time the server is busy servicing class i customers, and by B_i' the event that the server is not servicing queue i.

Theorem 2. In a polling system satisfying assumptions A.1–A.4, where the ith arrival process is MGE$_{M_i}$, characterized by matrices $A_{i,0}$ and $A_{i,1}$,

$$\boldsymbol{P}_{Q_i}(z) = \boldsymbol{P}_{Q_i|B_i'}(z)(1-\rho)(1-z)\bigl(\Phi_{X_i}(A_{i,0}+zA_{i,1}) - zI\bigr)^{-1}, \tag{3}$$

where $\boldsymbol{P}_{Q_i|B_i'}(z)$ is the vector generating function of the number of class i customers in the system at a random observation time when the server is *not* servicing class i customers.

Proof. From theorem 1 and the fact that $S_i = W_i + X_i$ and W_i, X_i are independent, we have

$$\begin{aligned}\boldsymbol{P}_{L_i}(z) &= \lambda_i(1-z)\,\boldsymbol{e}_1\Phi_{S_i}(A_{i,0}+zA_{i,1})(A_{i,0}+zA_{i,1})^{-1}\\ &= \lambda_i(1-z)\,\boldsymbol{e}_1\Phi_{X_i}(A_{i,0}+zA_{i,1})\Phi_{W_i}(A_{i,0}+zA_{i,1})(A_{i,0}+zA_{i,1})^{-1}.\end{aligned}$$

Combining the above equation with equation (2) we obtain that

$$\boldsymbol{P}_{L_i}(z) = \boldsymbol{P}_{Q_i}(z)\Phi_{X_i}(A_{i,0}+zA_{i,1}). \tag{4}$$

By applying Little's law to the server, $P\{B_i\} = \rho_i$, and $P\{B_i'\} = 1 - \rho_i$. By conditioning on the event B_i we obtain

$$\boldsymbol{P}_{L_i}(z) = z\boldsymbol{P}_{Q_i}(z) + (1 - z)(1 - \rho_i)\boldsymbol{P}_{Q_i|B_i'}(z), \tag{5}$$

where $\boldsymbol{P}_{Q_i|B_i'}(z)$ is the vector generating function of the number in the system from class i given that the server is *not* servicing that class. Combining equations (4) and (5) we prove equation (3). □

Remarks. 1. If we apply the above analysis to queue i in isolation, the event B_i' corresponds to the system being empty at a random observation time and, therefore, $\boldsymbol{P}_{Q_i|B_i'}(z) = (1 - \rho)^{-1}\boldsymbol{H}_i$, where $\boldsymbol{H}_i$ has the following meaning: $H_{i,j} \triangleq \Pr\{L_i = 0, R_i = j\}$ with R_i being the ATC of the ith arrival process at a random observation time. Hence, we obtain that the vector generating function of the number of customers waiting in the ith queue in isolation, $\boldsymbol{P}_{Q_i^{\mathrm{o}}}(z)$, is given by

$$\boldsymbol{P}_{Q_i^{\mathrm{o}}}(z) = \boldsymbol{H}_i(1 - z)\big(\Phi_{X_i}(A_{i,0} + zA_{i,1}) - zI\big)^{-1}. \tag{6}$$

This is in agreement with [3, proposition 1].

2. Notice that equation (3) can be equivalently written as

$$\boldsymbol{P}_{Q_i}(z) = \boldsymbol{P}_{Q_i|B_i'}(z)\Pi^{\mathrm{o}}(z),$$

where $\Pi^{\mathrm{o}}(z) \triangleq (1 - \rho)(1 - z)(\Phi_{X_i}(A_{i,0} + zA_{i,1}) - zI)^{-1}$ depends entirely on the characteristics of the ith queue. Therefore, theorem 2 has a decomposition character. This fact becomes more apparent in the case of Poisson arrivals where the vector generating functions become scalars. Then, equation (3) yields that the number of customers waiting in the ith queue of an arbitrary polling system decomposes into the number of customers waiting in the ith queue in isolation plus the number of customers waiting in the ith queue of the polling system when the server is on vacation from this queue. Therefore, it generalizes the results in [23], where the authors prove decomposition results for exhaustive and gated cyclic systems with Poisson arrivals, in that it allows for general polling and service policy and dependencies between the service times of the different queues.

To further illustrate the decomposition character of theorem 2 we state the following corollary.

Corollary 3. In a polling system satisfying assumptions A.1–A.4, where the ith arrival process is MGE_{M_i}, the expected number of customers waiting in queue decomposes as follows:

$$E[Q_i] = E\big[Q_i^{\mathrm{o}}\big] + E\big[Q_i \mid B_i'\big],$$

where $E[Q_i^{\mathrm{o}}]$ is the expected number of class i customers in queue for the $\mathrm{MGE}_{M_i}/\mathrm{G}/1$ system in isolation and $E[Q_i \mid B_i']$ is the expected number of class i customers in the system at a random observation time when the server is *not* servicing class i.

Theorem 2 holds under dynamic polling order and arbitrary service policies. We next establish sharper decomposition results for systems under static polices, satisfying assumptions A.1–A.7. For this class of systems we denote by $E = \{1, 2, \ldots, N\}$ the set of all queues. We let E_g, $E \setminus E_g$ be the set of queues served in a gated and exhaustive mode, respectively. We also let Δ_i^* be the age of the intervisit time and Λ_i^* be the elapsed time from the beginning of a cycle for queue i until a random observation time that occurs during the intervisit time Δ_i.

Theorem 4. In a polling system satisfying assumptions A.1–A.7, where the ith arrival process is MGE_{M_i}, characterized by matrices $A_{i,0}$ and $A_{i,1}$, we have that:

$$\boldsymbol{P}_{Q_i}(z) = \boldsymbol{P}_{Q_i^{\mathrm{o}}}(z)\Phi_{\Delta_i^*}(A_{i,0} + zA_{i,1}) \quad \text{and} \quad W_i \overset{d}{=} W_i^{\mathrm{o}} + \Delta_i^*, \quad i \in E \setminus E_g, \tag{7}$$

$$\boldsymbol{P}_{Q_i}(z) = \boldsymbol{P}_{Q_i^{\mathrm{o}}}(z)\Phi_{\Lambda_i^*}(A_{i,0} + zA_{i,1}) \quad \text{and} \quad W_i \overset{d}{=} W_i^{\mathrm{o}} + \Lambda_i^*, \quad i \in E_g, \tag{8}$$

where $\boldsymbol{P}_{Q_i^{\mathrm{o}}}(z)$ and W_i^{o} indicate the vector generating function of the number of class i customers in queue, and the waiting time, respectively, of an $\mathrm{MGE}_{M_i}/\mathrm{G}/1$ queue in isolation.

Proof. We first consider queue $i \in E \setminus E_g$. In this case the customers waiting in queue i at a random observation time during an intervisit period for queue i, must have arrived during the elapsed intervisit time for queue i. We denote by $R_{i,0}$ the ATC stage of the ith arrival process when the server leaves queue i and, therefore, he starts an intervisit interval with respect to queue i, Δ_i. We also denote by $\alpha_{i,k}(t)$ ($\alpha_{i,1}(t) = \alpha_i(t)$) the pdf of the remaining interarrival time for a class i customer in the kth stage of the ATC. Finally, we denote by $a_{i,r}^j(t)$ the probability for a class i customer in the ATC to move from stage $r \leqslant j$ to stage j during the interval $[0, t)$. Then we obtain, for $n \geqslant 1$,

$$\begin{aligned}
&P\bigl\{Q_i = n,\ R_i = j \mid B_i'\bigr\} \\
&\quad = \sum_{k=1}^{M_i} \int_0^\infty P\bigl\{Q_i = n,\ R_i = j \mid B_i',\ \Delta_i^* \in \mathrm{d}t,\ R_{i,0} = k\bigr\} P\bigl\{R_{i,0} = k,\ \Delta_i^* \in \mathrm{d}t\bigr\} \\
&\quad = \sum_{k=1}^{M_i} \int_0^\infty a_{i,k}(t) \cdot a_i^{(n-1)}(t) \cdot a_{i,1}^j(t) P\bigl\{R_{i,0} = k,\ \Delta_i^* \in \mathrm{d}t\bigr\}.
\end{aligned}$$

Similarly, for $n = 0$ we have that

$$P\bigl\{Q_i = 0, R_i = j \mid B_i'\bigr\} = \sum_{k=1}^{M_i} \int_0^\infty a_{i,k}^j(t) P\bigl\{R_{i,0} = k,\ \Delta_i^* \in \mathrm{d}t\bigr\}.$$

Taking z-transforms in the above relationship and using the fact (proven in [3]):

$$\mathrm{e}^{-(A_0+zA_1)t} = \begin{bmatrix} a_{i,1}^1(t) & \dots & a_{i,1}^M(t) \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{i,M}^M(t) \end{bmatrix} + \sum_{n=1}^{\infty} z^n \begin{bmatrix} a_{i,1}(t) \\ \vdots \\ a_{i,M}(t) \end{bmatrix} \cdot a_{i,1}^{(n-1)}(t) \cdot \bigl(a_{i,1}^1(t) \ \dots \ a_{i,1}^M(t)\bigr),$$

we obtain

$$\boldsymbol{P}_{Q_i|B_i'}(z) = \sum_{k=1}^{M_i} \int_0^\infty \boldsymbol{e}_k \mathrm{e}^{-(A_{i,0}+zA_{i,1})t} P\bigl\{R_{i,0} = k,\ \Delta_i^* \in \mathrm{d}t\bigr\}.$$

Assumptions A.5 and A.7 imply that Δ_i^* depends on the arrivals in all other queues but queue i. Therefore, since all arrival processes are independent (assumption A.4), the random variables $R_{i,0}$ and Δ_i^* are also independent. Hence,

$$\boldsymbol{P}_{Q_i|B_i'}(z) = \boldsymbol{R}_{i,0} \int_0^\infty \mathrm{e}^{-(A_{i,0}+zA_{i,1})t} P\bigl\{\Delta_i^* \in \mathrm{d}t\bigr\} = \boldsymbol{R}_{i,0} \Phi_{\Delta_i^*}(A_{i,0} + zA_{i,1}). \quad (9)$$

Taking limits as $z \to 1$, we have from equation (3) that

$$\boldsymbol{H}_i = (1-\rho_i) \lim_{z\to 1} \boldsymbol{P}_{Q_i|B_i'}(z) = (1-\rho_i)\boldsymbol{R}_{i,0},$$

which combined with equations (6) and (9) completes the proof of the first part of equation (7). The decomposition of the waiting times follows easily from the distributional law.

For the case of queue $i \in E_g$, the customers waiting in queue at a random observation time when the server is on vacation from queue i must have arrived during Λ_i^*. Following a line of arguments similar to the proof of equation (7), equation (8) follows. □

We next relate Λ_i^* to the characteristic quantities of the polling system, namely, Δ_i and T_i, as follows:

$$\phi_{\Lambda_i^*}(s) = E\bigl[\mathrm{e}^{-s(T_i+\Delta_i^*)}\bigr] = \int_0^\infty E\bigl[\mathrm{e}^{-s(T_i+\Delta_i^*)} \mid \Delta_i \in \mathrm{d}x,\ B_i'\bigr] P\bigl\{\Delta_i \in \mathrm{d}x \mid B_i'\bigr\}.$$

Conditioned on the duration of the intervisit interval the random variables T_i and Δ_i^* are independent and, hence,

$$\phi_{\Lambda_i^*}(s) = \int_0^\infty E\bigl[\mathrm{e}^{-sT_i} \mid \Delta_i \in \mathrm{d}x,\ B_i'\bigr] E\bigl[\mathrm{e}^{-s\Delta_i^*} \mid \Delta_i \in \mathrm{d}x,\ B_i'\bigr] \times P\bigl\{\Delta_i \in \mathrm{d}x \mid B_i'\bigr\}. \quad (10)$$

Due to the 'length biased' effect we have (see, for example, Baccelli and Bremaud [1])

$$E\big[\mathrm{e}^{-s\Delta_i^*} \mid x \leqslant \Delta_i < x + \mathrm{d}x,\ B_i'\big] = \int_{u=0}^{x} \frac{1}{x}\,\mathrm{e}^{-su}\,\mathrm{d}u.$$

Substituting into equation (10) we have that

$$\begin{aligned}\phi_{\Lambda_i^*}(s) &= \int_{x=0}^{\infty}\int_{t=0}^{\infty} \mathrm{e}^{-st} P\big\{T_i \in \mathrm{d}t,\ \Delta_i \in \mathrm{d}x \mid B_i'\big\}\left[\frac{1-\mathrm{e}^{-sx}}{sx}\right] \\ &= \frac{1}{sE[\Delta_i]}\int_{x=0}^{\infty}\int_{t=0}^{\infty} \mathrm{e}^{-st} P\big\{T_i \in \mathrm{d}t,\ \Delta_i \in \mathrm{d}x\big\}\big[1-\mathrm{e}^{-sx}\big], \end{aligned} \tag{11}$$

where we used the fact that

$$P\big\{T_i \in \mathrm{d}t,\ \Delta_i \in \mathrm{d}x \mid B_i'\big\} = \frac{x}{E[\Delta_i]} P\{T_i \in \mathrm{d}t,\ \Delta_i \in \mathrm{d}x\}.$$

Remarks. 1. The form of decomposition results in equations (7) and (8) does not depend on the specifics of the static polling policy and therefore it holds for cyclic and probabilistic routing policies as well as for policies with fixed-order polling tables.

2. Similar results hold under different *static* service policies, and can be proved using theorem 2 and evaluating $\boldsymbol{P}_{Q_i|B_i'}(z)$. For example, if the service policy is *reserved gated cyclic*, where the server serves exactly those class i customers that were present upon his departure from queue $i-1$, then

$$\boldsymbol{P}_{Q_i}(z) = \boldsymbol{P}_{Q_i^{\mathrm{o}}}(z)\boldsymbol{\Phi}_{\bar{\Lambda}_i^*}(A_{i,0} + zA_{i,1}),$$

where $\bar{\Lambda}_i^* \triangleq d_{i-1,i} + \Lambda_i^*$, since under this discipline the customers waiting in the ith queue at a random observation time while the server is *not* servicing this queue must have arrived either during the switch-over time from queue $i-1$ or during the elapsed time from the beginning of the cycle for queue i.

3. The expected waiting time decomposes as follows:

$$E[W_i] = \begin{cases} E\big[W_i^{\mathrm{o}}\big] + \dfrac{E[T_i\Delta_i]}{E[\Delta_i]} + \dfrac{E[(\Delta_i)^2]}{2E[\Delta_i]}, & i \in E_g, \\ E\big[W_i^{\mathrm{o}}\big] + \dfrac{E[(\Delta_i)^2]}{2E[\Delta_i]}, & i \in E \setminus E_g. \end{cases}$$

4. For systems with zero switch-over times, the above analysis carries on until the point we claimed that Δ_i^*, Λ_i^* are independent of the ith arrival process. This is not true anymore since the server starts idling when the system empties and then waits for the first arrival.

5. Theorem 4 generalizes the decomposition results in [23] in that it allows for more general arrival processes and arbitrary static polling policies. On the other hand, [23] also focuses on zero switch-over time systems, which are not discussed in this paper.

4. Decomposition results for polling systems in heavy traffic

In this section we consider the class of polling systems with general renewal arrival processes under heavy traffic conditions. Let us first define the term *'heavy traffic conditions'*. Intuitively, we say that a queue i is operating under heavy traffic conditions when both the number of customers and the waiting time are very large, namely, $P[Q_i < \infty,\ W_i < \infty] \to 0$. For a queue in isolation, in which either the interarrival or the service times are *nonarithmetic*, 'heavy traffic conditions' are equivalent to $\rho \to 1$. For a polling system, the analysis is more complicated and what we term 'heavy traffic conditions' depends on the service policy.

In particular, let π denote a combination of polling order and service policy. We fix the service time distribution and the switch-over times. Next, we define the set of arrival rates for which the polling system is stable, i.e., $\mathcal{L} \overset{\Delta}{=} \{\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)\colon\ \pi \text{ is stable}\}$. For example, in the case of the exhaustive, gated service policy or binomial-gated policy, it is well known that $\mathcal{L} = \{\boldsymbol{\lambda}\colon\ \rho < 1\}$ under minimal assumptions on the polling order (see [21,22]). However, for the 1-limited service policy the stability conditions are different and $\mathcal{L} = \{\boldsymbol{\lambda}\colon\ \lambda_i E[C_i] < 1\}$ for any Markovian polling order (see [9]). Note that $E[C_i]$ depends in general on the arrival rates, for example, in the case of cyclic polling order

$$E[C_i] = \frac{\sum_i E[d_{i,i+1}]}{1-\rho}.$$

Let $\mathcal{B}(\mathcal{L})$ be the boundary of $\mathcal{L}$.

We say that queue i is operating under heavy traffic conditions if either:

1. $\boldsymbol{\lambda} \to \boldsymbol{\lambda}_{\mathrm{o}} \in \mathcal{B}(\mathcal{L})$, given that as $\boldsymbol{\lambda} \to \boldsymbol{\lambda}_{\mathrm{o}}$, $P[Q_i < \infty, W_i < \infty] \to 0$.
2. $d_{ji} \to \infty$ for at least one j, given that as $d_{ji} \to \infty$, $P[Q_i < \infty, W_i < \infty] \to 0$.

Notice that as $\boldsymbol{\lambda} \to \boldsymbol{\lambda}_{\mathrm{o}} \in \mathcal{B}(\mathcal{L})$ at least one queue will be in heavy traffic conditions – but not necessarily all, as it may be the case with the k-limited policy. We use the expression $h(x) \sim r(x)$ to represent that $h(x)/r(x) = 1$, under heavy traffic conditions. Moreover, the expression $Y_1 \sim Y_2$, where Y_1 and Y_2 are random variables, is used as equivalent to $g_{Y_1}(x) \sim g_{Y_2}(x)$, where $g_{Y_i}(x)$ is the z-transform or the Laplace–Stieltjes transform (depending on whether we have discrete or continuous random variables) of Y_i. For systems in heavy traffic, we have the following result.

Theorem 5 (Bertsimas and Mourtzinou [3]). For a queue under heavy traffic conditions, that satisfies assumptions A.1–A.3, the distributional laws take the form:

$$E\left[z^{L_i}\right] \sim \phi_{S_i}\bigl(f_i(z)\bigr) \quad \text{and} \quad E\left[z^{Q_i}\right] \sim \phi_{W_i}\bigl(f_i(z)\bigr),$$

where $f_i(z) \overset{\Delta}{=} \lambda_i(1-z) - \frac{1}{2}\lambda_i(1-z)^2(c_{a_i}^2 - 1)$, with $c_{a_i}^2 < \infty$ being the squared coefficient of variation of the ith arrival process.

Using theorem 5 and following the same line of arguments with theorem 2, we prove the main decomposition result for polling systems in heavy traffic.

Theorem 6. In a polling system that satisfies assumptions A.1–A.4, the number of customers in the ith queue operating under heavy traffic conditions decomposes as follows:

$$E\bigl[z^{Q_i}\bigr] \sim E\bigl[z^{Q_i^{\mathrm{o}}}\bigr]E\bigl[z^{Q_i} \mid B_i'\bigr], \tag{12}$$

where $E[z^{Q_i^{\mathrm{o}}}]$ is the z-transform of the number of customers in a $GI/GI/1$ queue with the same arrival and service characteristics as queue i and $E[z^{Q_i} \mid B_i']$ is the z-transform of the number of customers in queue i given that the server is *not* servicing queue i.

If we further restrict the analysis to static polling order and gated or exhaustive polices, we obtain:

Theorem 7. In a polling system with renewal arrivals satisfying assumptions A.1–A.7, if queue i operates under heavy traffic conditions, then:

$$W_i \sim W_i^{\mathrm{o}} + \Lambda_i^* \quad \text{and} \quad Q_i \sim Q_i^{\mathrm{o}} + N_{a_i}^*\bigl(\Lambda_i^*\bigr), \quad i \in E_g, \tag{13}$$

$$W_i \sim W_i^{\mathrm{o}} + \Delta_i^* \quad \text{and} \quad Q_i \sim Q_i^{\mathrm{o}} + N_{a_i}^*\bigl(\Delta_i^*\bigr), \quad i \in E \setminus E_g, \tag{14}$$

where W_i^{o} and Q_i^{o} is the waiting time and the queue length in a regular $GI/GI/1$ queue, respectively, and Δ_i^* and Λ_i^* have been defined in section 3.

Proof. We first calculate $E[z^{Q_i} \mid B_i']$ if queue $i \in E_g$. Given the event B_i', the arrival of the random observer occurs during intervisit time Δ_i. Moreover, as the service policy is gated, the customers that are waiting in queue upon the arrival of the random observer must have arrived during the elapsed time from the beginning of the cycle C_i until the random observation time which occurred during Δ_i; which is denoted by Λ_i^*. Due to the heavy traffic assumptions we have therefore that

$$E\bigl[z^{Q_i} \mid B_i'\bigr] \sim E\bigl[z^{N_{a_i}^*(\Lambda_i^*)}\bigr] \sim \phi_{\Lambda_i^*}\bigl(f_i(z)\bigr), \tag{15}$$

where $\phi_{\Lambda_i^*}(s)$ has been calculated in equation (11). Combining equation (15) with equation (12) and the fact that from distributional laws in heavy traffic

$$E\bigl[z^{Q_i}\bigr] \sim \phi_{W_i}\bigl(f_i(z)\bigr) \quad \text{and} \quad E\bigl[z^{Q_i^{\mathrm{o}}}\bigr] \sim \phi_{W_i^{\mathrm{o}}}\bigl(f_i(z)\bigr),$$

we obtain the first part of equation (13). Similarly, we can prove the decomposition result for nodes served under exhaustive policy. □

The decomposition of the mean waiting times under heavy traffic conditions follows either by differentiating equations (13) and (14), or by differentiating the distributional laws (see [4] for a similar proof in the case of cyclic polling systems).

$$E[W_i] \sim \begin{cases} E\bigl[W_i^{\mathrm{o}}\bigr] + \dfrac{E[T_i\Delta_i]}{E[\Delta_i]} + \dfrac{E[(\Delta_i)^2]}{2E[\Delta_i]}, & i \in E_g, \\ E\bigl[W_i^{\mathrm{o}}\bigr] + \dfrac{E[(\Delta_i)^2]}{2E[\Delta_i]}, & i \in E \setminus E_g, \end{cases}$$

where

$$E\bigl[W_i^{\mathrm{o}}\bigr] \sim \frac{2\rho_i E[X_i^*] + E[X_i](c_{a_i}^2 - 1)}{2(1-\rho_i)}$$

is the mean waiting time in a regular $GI/GI/1$ queue and X_i^* is the age of the service time (see [3]).

5. Markovian polling systems under heavy traffic conditions

In this section we use the analytical results we have established so far to evaluate the performance of Markovian polling systems with gated service operating under heavy traffic conditions as defined in section 4. According to the Markovian scheme, the next station to be polled is determined from an irreducible Markov chain $\mathcal{M} = \{d_n,\ n = 0,1,\ldots\}$ with state space $\mathcal{I} = \{1,\ldots,N\}$. We denote by $\{e_n = i\}$ the event that the nth station to be polled after time $t = 0$ is station i, $i \in \mathcal{I}$. We assume that the Markov chain $\mathcal{M}$ has stationary one-step transition probabilities, i.e., the conditional probabilities $P\{e_{n+1} = j \mid e_n = i\}$ for all $i,j \in \mathcal{I}$ are independent of n. Then we define

$$p_{ij} \overset{\Delta}{=} P\{e_{n+1} = j \mid e_n = i\} \quad \text{and} \quad q_i \overset{\Delta}{=} \lim_{n\to\infty} P\{e_n = i\}, \quad i,j \in \mathcal{I},\ n = 0,1,\ldots.$$

We start our analysis by calculating the expected cycle and intervisit time for class i, $E[C_i]$ and $E[\Delta_i]$, via simple probabilistic arguments. We denote by ρ_i the traffic intensity of class i, by

$$\rho \overset{\Delta}{=} \sum_{i=1}^{N} \rho_i$$

the total traffic intensity, by

$$\sigma \overset{\Delta}{=} \sum_{i=1}^{N} q_i \sum_{j=1}^{N} p_{ij} d_{ij}$$

the average switch-over time, and by m_{ji} the average number of visits per unit time from queue i to queue j. From the above definitions we have that

$$E[C_i]\sum_{j=1}^{N} m_{ji} = 1 \quad \text{and} \quad E[\Delta_i]\sum_{j=1}^{N} m_{ij} = 1-\rho_i, \quad i = 1,\dots,N, \tag{16}$$

$$\frac{1-\rho}{\sigma} = \text{total number of switch-overs within a unit time} = \sum_{j=1}^{N}\sum_{i=1}^{N} m_{ji}. \tag{17}$$

Also from the definition of the steady-state probability q_i we have

$$q_i = \frac{\sum_{j=1}^{N} m_{ji}}{\sum_{j=1}^{N}\sum_{i=1}^{N} m_{ji}} = \frac{\sigma}{1-\rho}\sum_{j=1}^{N} m_{ji}, \quad i = 1,\dots,N,$$

where for the second equality we used equation (17). Combining the above relationship with equation (16) we get that

$$q_i E[C_i] = \frac{\sigma}{1-\rho} \quad \text{and} \quad E[\Delta_i] = (1-\rho_i)E[C_i], \quad i = 1,\dots,N. \tag{18}$$

We next evaluate $E[(\Delta_i)^2]$ and $E[T_i\Delta_i]$ and use the results of section 5 in order to obtain $E[W_i]$. We notice first that independent of the policy we follow, we always have that

$$\Delta_i^k = C_i^k - T_i^{k-1}. \tag{19}$$

Moreover, if we denote by N_i^k the number of customers that the server finds upon his arrival in the ith queue at his kth visit, we have from the definition of a *gated* policy that

$$T_i^k = \sum_{l=1}^{N_i^k} X_{i,l},$$

where T_i^k is the time the server spends servicing the ith queue in the kth visit and $X_{i,l}$ represents the service time distribution for the lth customer among N_i^k. Due to the nature of the gated policy, the N_i^k customers must have arrived during the cycle time C_i^k. Under heavy traffic conditions the intervisit time $\Delta_i^k \to \infty$ for all queues $i = 1,\dots,N$ and visits k. Hence, under heavy traffic conditions, the moment that the server enters queue i constitutes a random incidence for the ith arrival process. Therefore,

$$N_i^k \sim N_{a_i}^*\big(C_i^k\big) \quad \text{and} \quad T_i^k \sim \sum_{l=1}^{N_{a_i}^*(C_i^k)} X_{i,l}. \tag{20}$$

Based on equations (19) and (20) we will prove the following theorem.

Theorem 8. For a Markovian polling system in heavy traffic the mean waiting times $E[W_i]$ for all $i = 1, \dots, N$ are given as

$$E[W_i] \sim \frac{1+\rho_i}{2E[C_i]}\mathrm{var}[C_i] + \frac{(1+\rho_i)E[C_i]}{2} + \frac{(c_{a_i}^2 - 1)E[X_i]}{2}, \tag{21}$$

where the terms $\mathrm{var}[C_i]$ satisfy a linear system of equations (26)–(28).

Proof. Our strategy is to find $E[T_i\Delta_i]$ and $E[(\Delta_i)^2]$ as functions of $\mathrm{var}[C_i]$ and then form a linear system for $\mathrm{var}[C_i]$.

Step 1: Evaluation of $E[T_i\Delta_i]$ *and* $E[(\Delta_i)^2]$. Notice first that from equation (19) we have

$$\begin{aligned} E[T_i\Delta_i] &\overset{\Delta}{=} \lim_k E\big[T_i^{k-1}\Delta_i^k\big] = \lim_k E\big[T_i^{k-1}C_i^k\big] - \lim_k E\big[\big(T_i^{k-1}\big)^2\big] \\ &= \lim_k E\big[C_i^k T_i^{k-1}\big] - \lim_k \mathrm{var}\big[T_i^{k-1}\big] - \lim_k E\big[T_i^{k-1}\big]^2 \\ &= \beta_i + \rho_i(1-\rho_i)\big(E[C_i]\big)^2 - \mathrm{var}[T_i], \end{aligned} \tag{22}$$

where $\beta_i \overset{\Delta}{=} \lim_{k\to\infty} \mathrm{Cov}[C_i^k, T_i^{k-1}]$ and we used the fact that $E[T_i] = E[C_i] - E[\Delta_i] = \rho_i E[C_i]$.

On the other hand, we have from equation (19) that

$$\mathrm{var}\big[\Delta_i^k\big] = \mathrm{var}\big[C_i^k\big] + \mathrm{var}\big[T_i^{k-1}\big] - 2\mathrm{Cov}\big[C_i^k, T_i^{k-1}\big].$$

Taking limits as $k \to \infty$ and adding and subtracting $E[\Delta_i]^2$ we have that

$$E\big[(\Delta_i)^2\big] = \mathrm{var}[C_i] + \mathrm{var}[T_i] - 2\beta_i + (1-\rho_i)^2\big(E[C_i]\big)^2.$$

Next, we need to evaluate $\mathrm{var}[T_i]$. By differentiating equation (20) twice we obtain

$$E\big[(T_i)^2\big] \sim \rho_i^2 E\big[(C_i)^2\big] + \lambda_i E\big[(X_i)^2\big]E[C_i] + \lambda_i\big(c_{a_i}^2 - 1\big)\big(E[X_i]\big)^2 E[C_i].$$

We now use equation (18) and step 1 together with theorem 7 to obtain equation (21). To conclude the proof of the theorem we need to formulate the linear system that yields $\mathrm{var}[C_i]$ for all $i = 1, \dots, N$.

Step 2: Formulation of an $\mathrm{O}(N^3)$ *linear system.* We start by defining a reversed Markov chain of transitions obtained by $\mathcal{M}$ by inverting the time parameter. This reversed chain is stationary with the same equilibrium probabilities q_j, $j \in \mathcal{I}$, see [18], and one-step transition probabilities for $i, j \in \mathcal{I}$, given by

$$b_{ij} = P\{\text{the probability that the last stop before queue } i \text{ was queue } j\} = \frac{q_j}{q_i}p_{ji}.$$

Next, we decompose the interval between a random visit to node k and the previous visit to node i by conditioning on the node visited just before node k, and we obtain

$$C_{ki} = \begin{cases} C_{ji} + d_{jk} + T_j & \text{w.p. } b_{kj} \quad \text{for } j \neq i, \\ d_{ik} + T_i & \text{w.p. } b_{ki}. \end{cases} \tag{23}$$

By taking expectations we obtain the following $N^2 \times N^2$ system:

$$E[C_{ki}] = \sum_{j=1}^{N} b_{kj}\bigl(E[d_{jk}] + E[T_j]\bigr) + \sum_{\substack{j=1\\ j\neq k}}^{N} b_{kj} E[C_{ji}] \quad \text{for all } i, k.$$

Notice that we can decompose the above system for each i and obtain N independent $N \times N$ linear systems.

To proceed with our analysis we need to calculate var$[C_i] = E[C_{ii}C_{ii}] - (E[C_{ii}])^2$. From equation (23) we have that

$$\begin{aligned} E[C_{ii}C_{ii}] = {} & \sum_{j=1}^{N} b_{ij}\bigl(E\bigl[d_{ji}^2\bigr] + E\bigl[T_j^2\bigr] + 2E[d_{ji}]E[T_j]\bigr) \\ & + \sum_{\substack{j=1\\ j\neq i}}^{N} b_{ij}\bigl(2E[d_{ik}]E[C_{ji}] + E[C_{ji}C_{ji}] + 2E[T_jC_{ji}]\bigr). \end{aligned} \tag{24}$$

Since we have already expressed $E[(T_j)^2]$ as a function of $E[C_{jj}C_{jj}]$, only we need to find a relationship between $E[T_jC_{ji}]$, $E[C_{ji}C_{ji}]$ and other known quantities for all i, j where $i \neq j$. Using (23) we have that

$$\begin{aligned} E[T_jC_{ji}] = {} & \sum_{m=1}^{N} b_{jm} E\bigl[(d_{mj}T_j + T_mT_j) \mid m \text{ is visited just before } j\bigr] \\ & + \sum_{\substack{m=1\\ m\neq i}}^{N} b_{jm} E[T_jC_{mi} \mid m \text{ is visited just before } j]. \end{aligned} \tag{25}$$

From the defining relationships of the system it is straightforward to get

$$\begin{aligned} & E[T_mT_j \mid m \neq j \text{ is visited just before } j] \\ & \quad \sim E\left[T_m \sum_{l=1}^{N^*_{a_j}(T_m + d_{mj} + C_{mj})} X_{j,l}\right] = \rho_j\bigl(E\bigl[T_m^2\bigr] + E[d_{mj}]E[T_m] + E[T_mC_{mj}]\bigr), \end{aligned}$$

$$E[T_jT_j \mid j \text{ is visited just before } j] \sim \rho_j\bigl(E\bigl[T_j^2\bigr] + E[d_{jj}]E[T_j]\bigr),$$

$$\begin{aligned} & E[T_jC_{mi} \mid m \neq j,\ i \text{ is visited just before } j] \\ & \quad \sim \rho_j\bigl(E[T_mC_{mi}] + E[d_{mj}]E[C_{mi}] + E[C_{mi}C_{mj}]\bigr), \end{aligned}$$

$$E[T_jC_{ji} \mid j \neq i \text{ is visited just before } j] \sim \rho_j\big(E[T_jC_{ji}] + E[d_{jj}]E[C_{ji}]\big),$$
$$E[T_jd_{mj} \mid m \text{ is visited just before } j]$$
$$\sim \rho_j\big(E\big[(d_{mj})^2\big] + E[d_{mj}]E[C_{mj}] + E[d_{mj}]E[T_m]\big),$$
$$E[T_jd_{jj} \mid j \text{ is visited just before } j] \sim \rho_j\big(E\big[(d_{jj})^2\big] + E[d_{jj}]E[T_j]\big).$$

Substituting the above set of equations into equation (25) we obtain the following $N^2 \times N^2$ system:

$$\begin{aligned} E[T_jC_{ji}] = {}& \sum_{m=1}^{N} b_{jm}\rho_j\big(E\big[d_{mj}^2\big] + 2E[d_{mj}]E[T_m] + E\big[T_m^2\big]\big) \\ & + \sum_{\substack{m=1\\ m\neq i}}^{N} b_{jm}\rho_j\big(E[C_{mi}]E[d_{mj}] + E[T_mC_{mi}]\big) + \sum_{\substack{m=1\\ m\neq j,i}}^{N} b_{jm}\rho_j E[C_{mj}C_{mi}] \\ & + \sum_{\substack{m=1\\ m\neq j}}^{N} b_{jm}\rho_j\big(E[C_{mj}]E[d_{mj}] + E[T_mC_{mj}]\big). \end{aligned} \tag{26}$$

As for $E[C_{ji}C_{ji}]$ we can take second moments in equation (23) and obtain $E[C_{ki}C_{kr}]$ for all k, i, r. Recall that from equation (23) we have that

$$C_{ki} = \begin{cases} C_{ji} + d_{jk} + T_j & \text{w.p. } b_{kj} \quad \text{for } j \neq i, \\ d_{ik} + T_i & \text{w.p. } b_{ki} \end{cases}$$

and similarly

$$C_{kr} = \begin{cases} C_{jr} + d_{jk} + T_j & \text{w.p. } b_{kj} \quad \text{for } j \neq r, \\ d_{rk} + T_r & \text{w.p. } b_{kr}. \end{cases}$$

Assuming that $r \neq i$ we have that

$$\begin{aligned} E[C_{ki}C_{kr}] = {}& b_{ki}E\big[(d_{ik} + T_i)(d_{ik} + T_i + C_{ir})\big] + b_{kr}E\big[(d_{rk} + T_r + C_{ri})(d_{rk} + T_r)\big] \\ & + \sum_{j=1,\ j\neq i,r}^{N} b_{kj}E\big[(d_{jk} + T_j + C_{ji})(d_{jk} + T_j + C_{jr})\big]. \end{aligned}$$

Equivalently, we have

$$\begin{aligned} & E[C_{ki}C_{kr}] \\ & \quad = \sum_{j=1}^{N} b_{kj}\big(E\big[d_{jk}^2\big] + E\big[T_j^2\big] + 2E[d_{jk}]E[T_j]\big) \\ & \qquad + \sum_{\substack{j=1\\ j\neq i,r}}^{N} b_{kj}\big(E[d_{jk}]\big(E[C_{jr}] + E[C_{ji}]\big) + E[C_{ji}C_{jr}] + E[T_jC_{jr}] + E[T_jC_{ji}]\big) \end{aligned}$$

$$+\,b_{ki}\bigl(E[d_{ik}]E[C_{ir}]+E[T_iC_{ir}]\bigr)+b_{kr}\bigl(E[d_{rk}]E[C_{ri}]+E[T_rC_{ri}]\bigr). \qquad (27)$$

We can similarly get an expression for $E[C_{ki}C_{ki}]$, as follows:

$$\begin{aligned} E[C_{ki}C_{ki}] = {} & \sum_{j=1}^{N} b_{kj}\bigl(E\bigl[d_{jk}^2\bigr]+E\bigl[T_j^2\bigr]+2E[d_{jk}]E[T_j]\bigr) \\ & + \sum_{\substack{j=1\\ j\neq i}}^{N} b_{kj}\bigl(2E[d_{jk}]E[C_{ji}]+E[C_{ji}C_{ji}]+2E[T_jC_{ji}]\bigr). \end{aligned} \qquad (28)$$

Combining the last two equations with equation (26) results into an $(N^3+N^2)\times(N^3+N^2)$ system with unknowns the $E[C_{ki}C_{ri}]$ and $E[T_iC_{ki}]$, for all i,k,r. □

Remarks. 1. It is important to notice that we obtained equation (21) without using the fact that the polling policy is Markovian. In particular, we proved equation (21) for *any* static gated polling system with independent renewal arrival processes, under heavy traffic. For Poisson arrivals equation (21) is known to hold for cyclic systems (see [24]), for random polling (see [19]) and for systems with a fixed-order polling table (see [2]). A similar relationship can be proved, when the arrival processes are MGE.

2. In the special case where the polling policy is cyclic, our analysis is greatly simplified and we obtain the results of [4]. In the case where the polling policy is random, i.e., $p_{ij}=p_j$ for all $i,j=1,\dots,N$ and $d_{ij}=d_i$ we obtain a generalization of the results of Kleinrock and Levy [19] in the sense that we allow for general arrival processes.

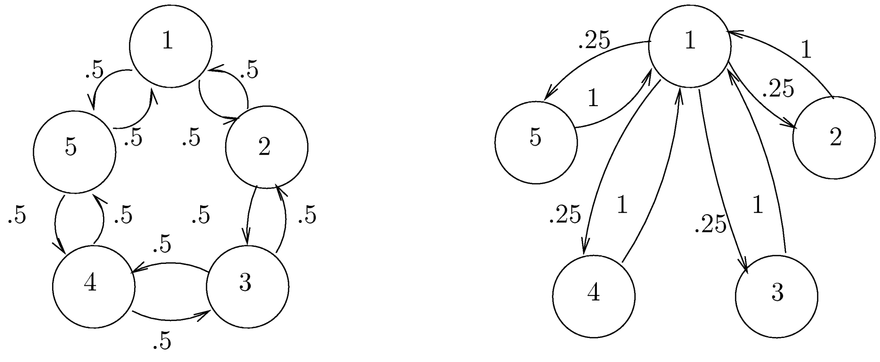
3. If the Markovian system is *symmetric*, i.e., the arrival and service processes and the switch-over time distributions are the same for all queues and also $p_{ij}=1/N$ for all $i,j=1,\dots,N$, we obtain closed-form expressions for the expected waiting times:

$$E[W_i]\sim\frac{(N-1)E[d]}{2(1-N\rho)}+\frac{N(1+\rho)E[d]}{2(1-N\rho)}+\frac{N\lambda E[X^2]}{2(1-N\rho)}+\frac{\mathrm{var}[d]}{2E[d]}+\frac{(c_a^2-1)E[X^2]}{2},$$

where λ and c_a^2 are the arrival rate and the square coefficient of arrival process, respectively, $E[X^2]$ is the second moment of the service time distribution, $\rho\overset{\Delta}{=}\lambda E[X]$ is the traffic intensity of the individual node and d is the switch-over time.

Numerical results

We now evaluate numerically our results for a number of 5-node Markovian systems under gated policy. First, we consider a series of symmetric systems with transition matrix described in figure 1(a). Then, we consider a series of asymmetric systems with transition matrix described in figure 1(b).



(a) Symmetric System

(b) Asymmetric System

Figure 1. Transition chains.

Table 1
Numerical results for 5-node symmetric polling systems.

ρ	$d = 0.5$			$d = 2.0$		
	$E[W^H]$	$E[W^S]$	Dev	$E[W^H]$	$E[W^S]$	Dev
0.5	5.250	5.323 (±0.113)	−1.371%	19.500	19.510 (±0.293)	−0.051%
0.6	6.875	6.927 (±0.197)	−0.751%	24.875	24.879 (±0.255)	−0.016%
0.7	9.583	9.618 (±0.113)	−0.364%	33.834	33.838 (±0.288)	−0.012%
0.8	15.000	15.020 (±0.294)	−0.133%	51.750	51.753 (±0.549)	−0.006%
0.9	31.250	31.258 (±0.617)	−0.025%	105.50	105.517 (±0.928)	−0.016%

For all systems, the arrival processes are Erlang 2 with $\lambda_i = \frac{1}{5}\rho$, and the services are Exponential with $\mu_i = 1$.

Table 1 presents the performance of our method, i.e., the accuracy of the average waiting time calculated via our heavy traffic method compared to the actual average waiting time obtained via simulation, as a function of the total traffic intensity, ρ, and the switch-over time, d, for the symmetric polling system and table 2 presents the corresponding results for the asymmetric system. We denote by $E[W_i^H]$ the waiting time for class i as obtained using theorem 7, by

$$E\left[W^H\right] \triangleq \sum_i \frac{\lambda_i}{\lambda} E\left[W_i^H\right],$$

and by $E[W^S]$ the average waiting time obtained via simulation. We also denote by Dev the deviation between the two methods defined as

$$\text{Dev} \triangleq \frac{E[W^H] - E[W^S]}{E[W^S]}.$$

Finally, we report the standard error of the simulation.

Table 2
Numerical results for 5-node asymmetric polling systems.

ρ	$d = 0.5$			$d = 2.0$		
	$E[W^H]$	$E[W^S]$	Dev	$E[W^H]$	$E[W^S]$	Dev
0.5	6.640	6.701 (±0.115)	−0.910%	25.060	25.066 (±0.206)	−0.024%
0.6	8.635	8.677 (±0.096)	−0.484%	31.915	31.917 (±0.216)	−0.006%
0.7	11.960	11.988 (±0.123)	−0.234%	43.340	43.340 (±0.969)	0.000%
0.8	18.610	18.626 (±0.299)	−0.086%	66.190	66.188 (±0.777)	0.015%
0.9	38.560	38.575 (±0.398)	−0.039%	134.740	134.745 (±1.486)	−0.037%

Table 3
Numerical results for 5-node asymmetric polling systems with $d = 0.5$.

ρ	$E[W_1^H]$	$E[W_1^S]$	Dev$_1$	$E[W_2^H]$	$E[W_2^S]$	Dev$_2$
0.5	1.447	1.563 (±0.017)	−7.421%	7.938	7.985 (±0.139)	−0.589%
0.6	2.000	2.085 (±0.018)	−4.077%	10.293	10.325 (±0.114)	−0.310%
0.7	2.897	2.954 (±0.026)	−1.930%	14.226	14.247 (±0.147)	−0.147%
0.8	4.656	4.689 (±0.069)	−0.704%	22.098	22.109 (±0.357)	−0.050%
0.9	9.862	9.877 (±0.098)	−0.152%	45.735	45.747 (±0.473)	−0.026%

Table 4
Numerical results for 5-node asymmetric polling systems with $d = 2.0$.

ρ	$E[W_1^H]$	$E[W_1^S]$	Dev$_1$	$E[W_2^H]$	$E[W_2^S]$	Dev$_2$
0.5	5.102	5.119 (±0.024)	−0.332%	30.049	30.051 (±0.157)	−0.007%
0.6	6.838	6.845 (±0.033)	−0.102%	33.184	38.187 (±0.196)	−0.009%
0.7	9.756	9.759 (±0.155)	−0.031%	51.736	51.738 (±0.857)	−0.004%
0.8	15.632	15.632 (±0.145)	0.000%	78.830	78.831 (±0.752)	−0.001%
0.9	33.341	33.342 (±0.291)	−0.003%	160.090	160.096 (±1.414)	−0.004%

Next, in tables 3 and 4 we compare our results for the individual waiting times for the system of figure 1(b), with switch-over time $d = 0.5$ and $d = 2.0$, respectively. Due to the topology of the system, classes 2–5 have the same expected waiting time. Hence, we report the expected waiting time for classes 1 and 2, $E[W_1^H]$ and $E[W_2^H]$, respectively. Similarly we report $E[W_1^S]$ and $E[W_2^S]$, the expected waiting time for classes 1 and 2 obtained via simulation. We also denote by Dev$_i$ the deviation between the two methods defined as

$$\mathrm{Dev}_i \triangleq \frac{E[W_i^H] - E[W_i^S]}{E[W_i^S]} \quad \text{for } i = 1, 2.$$

Finally, we report the standard error of the simulation.

As expected, for any particular system the performance of our method improves as the traffic intensity increases. Moreover, by comparing the accuracy of our methods for different systems, we see that our method performs better as the waiting time increases. For example, for traffic intensity $\rho = 0.5$ our prediction for the expected waiting time is very accurate, in both symmetric and asymmetric systems with switch-over times $d = 2.0$, but not as accurate for the symmetric system with switch-over times $d = 0.5$. Similarly, it is not as accurate for the mean waiting time for node 1 (the node with the smallest wait) in the asymmetric case as it is for the mean waiting time of the other nodes. These remarks are consistent with the conclusions drawn from numerical results in the cases of single class and multiclass priority queues in [4] and are based on the nature of our asymptotic method.

6. Concluding remarks

In this paper we established a number of structural relationships for polling systems. We started our analysis allowing for dynamic polling policies and demonstrated that the number of customers in each node of the system consists of one component that depends entirely on the specific characteristics of this node in isolation and another component that incorporates the dependencies introduced by the polling setting. By considering static polling policies (assumptions A.5–A.7), we further characterized the second component and obtained sharper decomposition results for both the number of customers in the system and the waiting time of each customer class.

The derived decomposition results apart from enhancing our understanding of the polling systems mechanism, can be used to obtain the performance analysis of specific systems as we illustrated in section 5. Moreover, since our methodology was based on distributional laws, our results further demonstrate the importance of using the unified approach we proposed in [3] to address various queueing problems.

References

[1] F. Baccelli and P. Bremaud, *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences* (Springer, New York, 1994).

[2] J.E. Baker and I. Rubin, Polling with a general-service order table, IEEE Trans. Commun. 35 (1987) 283–288.

[3] D. Bertsimas and G. Mourtzinou, A unified method to analyze overtake-free queueing systems, Adv. in Appl. Probab. 28 (1996) 588–625.

[4] D. Bertsimas and G. Mourtzinou, Multiclass queueing systems in heavy traffic: An asymptotic approach based on distributional and conservation laws, Oper. Res. 45(3) (1997) 470–487.

[5] D. Bertsimas and D. Nakazato, The distributional Little's law and its applications, Oper. Res. 43 (1995) 298–310.

[6] S.C. Borst and O.J. Boxma, Polling models with and without switchover times, Oper. Res. 45 (1997) 536–543.

[7] O.J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, Queueing Systems 5 (1989) 185–214.

[8] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, J. Appl. Probab. 24 (1987) 949–964.

[9] O.J. Boxma and J.A. Weststrate, Waiting times in polling systems with Markovian server routing, in: *Lecture Notes in Computer Science* 218, eds. G. Stiege and J.S. Lie (Academic Press, Berlin, 1989) pp. 89–104.

[10] R.B. Cooper and G. Murray, Queues served in cyclic order, Bell Syst. Tech. J. 48 (1969) 675–689.

[11] D.R. Cox, The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables, Proc. Cambridge Philos. Soc. 51 (1955) 433–441.

[12] M. Eisenberg, Queues with periodic service and change-over times, Oper. Res. 20 (1972) 440–451.

[13] S.W. Fuhrmann, Symmetric queues served in cyclic order, Oper. Res. Lett. 4 (1985) 139–144.

[14] S.W Fuhrmann and R.B. Cooper, Stochastic decompositions in a $M/G/1$ queue with generalized vacation, Oper. Res. 33 (1985) 1117–1129.

[15] R. Haji and G. Newell, A relation between stationary queue and waiting time distributions, J. Appl. Probab. 8 (1971) 617–620.

[16] J. Keilson and L. Servi, A distributional form of Little's law, Oper. Res. Lett. 7 (1988) 223–227.

[17] J. Keilson and L. Servi, The distributional form of Little's law and the Fuhrmann–Cooper decomposition, Oper. Res. Lett. 9 (1990) 239–247.

[18] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).

[19] L. Kleinrock and H. Levy, The analysis of random polling systems, Oper. Res. 36 (1988) 716–732.

[20] A.G. Konheim and B. Meister, Waiting lines and times in a system with polling, J. Assoc. Comput. Mach. 21 (1974) 470–490.

[21] H. Levy, Binomial-gated service: A method for effective operation and optimization of polling systems, IEEE Trans. Commun. 39 (1991) 1341–1349.

[22] H. Levy and M. Sidi, Polling systems: Applications, modeling and optimization, IEEE Trans. Commun. 38 (1990) 1750–1760.

[23] S.C. Srinivasan, M.M. Niu and R.B. Cooper, Relating polling systems with zero and nonzero switchover times, Queueing Systems 19 (1995) 149–168.

[24] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).

[25] H. Takagi, ed., *Stochastic Analysis of Computer and Communication Systems* (Elsevier, Amsterdam, 1990).

[26] P. Tran-Gia, Analysis of polling systems with general input process and finite capacity, IEEE Trans. Commun. 40 (1992) 337–344.