Towards a Unifying Theory on Branching-type Polling Systems in Heavy Traffic

R.D. van der Mei[a,b]

[a]Centre for Mathematics and Computer Science, Amsterdam, Netherlands
[b]Vrije Universiteit, Mathematics and Computer Science, Amsterdam, Netherlands

For a broad class of polling models the evolution of the system at specific embedded polling instants is known to constitute multi-type branching process (MTBP) with immigration. In this paper it is shown that for this class of polling models the vector that describes the state of the system at these polling instants, say $\underline{X} = (X_1, \ldots, X_M)$, satisfies the following heavy-traffic behavior (under mild assumptions):

$$(1-\rho)\underline{X} \to_d \underline{\gamma}\, \Gamma(\alpha,\mu) \quad (\rho \uparrow 1), \tag{1}$$

where $\underline{\gamma}$ is a known M-dimensional vector, $\Gamma(\alpha,\mu)$ has a gamma-distribution with known parameters α and μ, and where ρ is the load of the system. This general and powerful result is shown to lead to exact - and in many cases even closed-form - expressions for the Laplace-Stieltjes Transform (LST) of the complete asymptotic queue-length and waiting-time distributions for a broad class of branching-type polling models that includes many well-studied polling models policies as special cases. The results generalize and unify many known results on the waiting times in polling systems in heavy traffic, and moreover, lead to new exact results for classical polling models that have not been observed before. To demonstrate the usefulness of the results, we derive closed-form expressions for the LST of the waiting-time distributions for models with cyclic globally-gated polling regimes, and for cyclic polling models with general branching-type service policies. As a by-product, our results lead to a number of asymptotic insensitivity properties, providing new fundamental insights in the behavior of polling models.

Keywords: Polling systems, multi-type branching processes, heavy traffic, waiting-time distribution, queue-length distribution, gamma-distribution, unifying theory

1 Introduction

Polling systems are multi-queue systems in which a single server visits the queues in some order to serve the customers waiting at the queues, typically incurring some amount of switch-over time to proceed from one queue to the next. Polling models find a wide variety of applications in which processing power (e.g., CPU, bandwidth, manpower) is shared among different types of users. Typical application areas of polling models are computer-communication systems, logistics, flexible manufacturing systems, production systems and maintenance systems; the reader is referred to [35, 20] for extensive overviews of the applicability of polling models. Over the past few decades the performance analysis of polling models has received much attention in the literature. We refer to the classical surveys by [33, 34], and to a recent survey paper by Vishnevskii and Semenova [46] for overviews of the available results on polling models. One of the most remarkable results is that there appears to be a striking difference in complexity between polling models. Resing [29] observed that for a large class of polling models, including for example cyclic polling models with Poisson arrivals and exhaustive and gated service at all queues, the evolution of the system at successive polling instants at a fixed queue can be described as a multi-type branching process (MTBP) with immigration. Models that satisfy this MTBP-structure allow for an exact analysis, whereas models that violate the MTBP-structure are often more intricate.

In this paper we study the heavy-traffic behavior for the class of polling models that have an MTBP-structure, in a general parameter setting. Initiated by the pioneering work of Coffman et al. [11, 12], the analysis of the heavy-traffic behavior of polling models has gained a lot of interest over the past decade. This has led to the derivation of asymptotic expressions for key performance metrics, such as the moments and distributions of the waiting times and the queue lengths, for a variety of model variants, including for example models with mixtures of exhaustive and gated service policies with cyclic server routing [36], periodic server routing [43, 44], simultaneous batch arrivals [39], continuous polling [17], amongst others. In this context, a remarkable observation is that in the heavy-traffic behavior of polling models a central role

limiting distribution of the (scaled) cycle times and the marginal queue-lengths at polling instants. This observation has motivated us to develop a unifying theory on the heavy-traffic behavior of polling models that includes all these model instances as special cases, where everything falls into place. We believe that the results presented in this paper are a significant step towards such a general unifying theory.

The motivation for studying heavy-traffic asymptotics in polling models is twofold. First, a particularly attractive feature of heavy-traffic asymptotics (i.e., when the load tends to 1) for MTBP-type models is that in many cases they lead to strikingly simple expressions for queue-length and waiting-time distributions, especially when compared to their counterparts for arbitrary values of the load, which usually leads to very cumbersome expressions, even for the first few moments (cf., e.g., [18]). The remarkable simplicity of the heavy-traffic asymptotics provides fundamental insight in the impact of the system parameters on the performance of the system, and in many cases attractive insensitivity properties have been observed (see also Sections 3.1 and 3.2). A second motivation for considering heavy-traffic asymptotics is that the computation time needed to calculate the relevant performance metrics usually become prohibitively long when the system is close to saturation, both for branching-type [10] and non-branching-type polling models [3, 4], which raises the need for simple and fast approximations. To this end, heavy-traffic asymptotics form an excellent basis for developing such approximations (see also Section 3.3), and in fact, have been found to be remarkably accurate in several cases, even for moderate load (cf., e.g., [36, 38, 44]).

Recently, polling models in heavy traffic have received attention in the literature, and significant progress has been made in this area. For a two-queue model with exhaustive service and independent renewal arrival processes, Coffman et al. [11, 12] use the theory of diffusion processes to derive expressions for the joint workload distribution and the waiting-time distributions under heavy traffic assumptions. For models with independent Poisson arrivals, Kudoh et al. [18] give explicit expressions for the second moment of the waiting time in fully symmetric systems with gated or exhaustive service at each queue for models with two, three and four queues, by exploring the classical buffer-occupancy approach (cf., e.g., [33]), which is based on the relation between the joint queue-length distributions at successive polling instants. They also give conjectures for the heavy-traffic limits of the first two moments of the waiting times for systems with an arbitrary number of queues. In a series of papers, Van der Mei and co-authors explore the use of the Descendant Set Approach (DSA) [16] to derive exact expressions the waiting-time distributions in models with mixtures of exhaustive and gated service and cyclic [36] or periodic [43] server routing. Following a similar approach, Van der Mei also derives the exact asymptotics waiting-time distribution in cyclic queueing models with simultaneous batch arrivals [39]. Kroese [17] studies continuous polling systems in heavy traffic with unit Poisson arrivals on a ring and shows that the steady-state number of customers at each queue has approximately a gamma-distribution. Vatutin and Dyakonova [45] use the theory of MTBPs to obtain the limiting distributions several two-queue polling models with zero switch-over times. In addition to the evaluation of the performance of heavily loaded polling systems, the results can also be used to address stochastic scheduling problems, see for example [22, 23, 27, 28] and referenes therein.

To develop a unifying theory on the heavy-traffic behavior of branching-type polling models, it is interesting to observe that the theory of MTBPs, which was developed largely developed in the early 1970s, is well-matured and powerful (cf., e.g., [26, 14, 15]). Nonetheless, the theory of MTBPs has received remarkably little attention in the literature on polling models. In fact, throughout this paper we will show that the following result on MTBPs can be used as the basis for the development of a unifying theory on branching-type polling models under heavy-traffic assumptions: the joint probability distribution of the M-dimensional branching process $\{\underline{Z}_n, n = 0, 1, \ldots\}$ (with immigration in each state) converges in distribution to $\underline{v}\Gamma(\alpha, \mu)$ in the sense that (cf. Quine [26]):

$$\lim_{n\to\infty} \frac{1}{\pi_n(\xi)}\, \underline{Z}_n \to_d \underline{v}\Gamma(\alpha, \mu) \qquad (\xi \uparrow 1), \tag{2}$$

where ξ is the maximum eigenvalue of the so-called mean matrix, $\pi_n(\xi)$ is a scaling function, $\underline{v}$ is a known M-dimensional vector and $\Gamma(\alpha, \mu)$ is a gamma-distributed random variable with known shape and scale parameters α and μ, respectively. We emphasize that (2) is valid for general MTBPs under very mild moment conditions (see Section 2 for details). In this paper, we show that this result (2) can be transformed into equation (1), providing an asymptotic analysis for a very general class of MTBP-type polling models. Subsequently, we show that equation (1) leads to exact asymptotic expressions for the scaled time-average queue-length and waiting-time distributions under heavy-traffic assumptions; for specific model instances, basically all we have to do is calculate the parameters $\underline{v}$, α and μ, and the derivative of ξ as a function of ρ at $\rho = 1$, which is usually straightforward. In this way, we propose a new and powerful approach to derive

results we use the approach developed in this paper to derive new and yet unknown closed-form expressions for the complete asymptotic waiting-time distributions for a number of classical polling models. To this end, we derive closed-form expressions for the asymptotic waiting-time distributions for cyclic polling models with the Globally-Gated (GG) service policy, and for models with general branching-type service policies. As a by-product, the results also lead to asymptotic insensitivity properties providing new fundamental insights in the behavior of polling models. Moreover, the results lead to simple approximatons for the waiting-time distributions in stable polling systems.

The remainder of this paper is organized as follows. In Section 2 we give a brief introduction on MTBPs and formulate the limiting result by Quine [26] (see Theorem 1) that will be used throughout. In Section 3 we translate this result to the context of polling models, and give an approach for how to obtain heavy-traffic asymptotics for branching-type polling models. To illustrate the usefulness of the approach, we consider two specific types of polling models: (1) cyclic models with GG service, and (2) cyclic models with general branching-type service policies. For these models, we derive a complete characterization of the asymptotic waiting-time distributions. The implications of these results are discussed extensively. Finally, in Section 4 we address a number of challenging topics for further research.

2 Multitype branching processes with immigration

We consider a general M-dimensional multi-type branching process $\mathbf{Z} = \{\underline{Z}_n, n = 0, 1, \ldots\}$, where $\underline{Z}_n = (Z_n^{(1)}, \ldots, Z_n^{(M)})$ is an M-dimensional vector denoting the state of the process in the n-th generation, and where $Z_n^{(i)}$ is the number of type-i particles in the n-th generation, for $i = 1, \ldots, M$, $n = 0, 1, \ldots$. The process $\mathbf{Z}$ is completely characterized by (1) its one-step offspring function and (2) its immigration function, which are assumed mutually independent and to be stochastically the same for each generation. The one-step offspring function is denoted by $f(\underline{z}) = (f^{(1)}(\underline{z}), \ldots, f^{(M)}(\underline{z}))$, with $\underline{z} = (z_1, \ldots, z_M)$, and where for $|z_k| \leq 1$ $(k = 1, \ldots, M), i = 1, \ldots, M$,

$$f^{(i)}(\underline{z}) = \sum_{j_1,\ldots,j_M \geq 0} p^{(i)}(j_1, \ldots, j_M) z_1^{j_1} \cdots z_M^{j_M}, \tag{3}$$

where $p^{(i)}(j_1, \ldots, j_M)$ is the probability that a type-i particle produces j_k particles of type k $(k = 1, \ldots, M)$. The immigration function is denoted as follows: For $|z_k| \leq 1$ $(k = 1, \ldots, M)$,

$$g(\underline{z}) = \sum_{j_1,\ldots,j_M \geq 0} q(j_1, \ldots, j_M) z_1^{j_1} \cdots z_M^{j_M}, \tag{4}$$

where $q(j_1, \ldots, j_M)$ is the probability that a group of immigrant consists of j_k particles of type k $(k = 1, \ldots, M)$. Denote

$$\underline{g} := (g_1, \ldots, g_M), \text{ where } g_i := \frac{\partial g(\underline{z})}{\partial z_i}|_{\underline{z}=\underline{1}}, \tag{5}$$

and where $\underline{1}$ is the M-vector where each component is equal to 1. A key role in the analysis will be played by the first and second-order derivatives of $f(\underline{z})$. The first-order derivatives are denoted by the mean matrix

$$\mathbf{M} = (m_{i,j}), \quad \text{with } m_{i,j} := \frac{\partial f^{(i)}(\underline{z})}{\partial z_j}|_{\underline{z}=\underline{1}} \quad (i, j = 1, \ldots, M). \tag{6}$$

Thus, adopting the standard notion of "children", for a given type-i particle in the n-th generation, $m_{i,j}$ is the mean number of type-j children it has in the $(n+1)$-st generation. Similarly, for a type-i particle, the second-order derivatives are denoted by the matrix

$$\mathbf{K}^{(i)} = \left(k_{j,k}^{(i)}\right), \quad \text{with } k_{j,k}^{(i)} := \frac{\partial^2 f^{(i)}(\underline{z})}{\partial z_j \partial z_k}|_{\underline{z}=\underline{1}}, \quad i, j, k = 1, \ldots, M. \tag{7}$$

Denote by $\underline{v} = (v_1, \ldots, v_M)$ and $\underline{w} = (w_1, \ldots, w_M)$ the left and right eigenvectors corresponding to the largest real-valued, positive eigenvalue ξ of $\mathbf{M}$, commonly referred to as the maximum eigenvalue (cf., e.g., [2]), normalized such that

$$\underline{v}^\top \underline{1} = \underline{v}^\top \underline{w} = 1. \tag{8}$$

$\xi < 1$ and

$$\sum_{j_1+\cdots+j_M>0} q(j_1,\ldots,j_M)\log(j_1+\cdots+j_M) < \infty. \tag{9}$$

Throughout the following definitions are convenient. For any variable x that depends on ξ we use the hat-notation $\hat{x}$ to indicate that x is evaluated at $\xi = 1$. Moreover, for $\xi \geq 0$ let

$$\pi_0(\xi) := 0, \quad \text{and} \quad \pi_n(\xi) := \sum_{r=1}^{n} \xi^{r-2}, \quad n = 1, 2, \ldots. \tag{10}$$

A non-negative continuous random variable $\Gamma(\alpha, \mu)$ is said to have a gamma-distribution with shape parameter $\alpha > 0$ and scale parameter $\mu > 0$ if it has the probability density function

$$f_\Gamma(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-\mu x} \quad (x > 0) \quad \text{with} \quad \Gamma(\alpha) := \int_{t=0}^{\infty} t^{\alpha-1} e^{-t} dt, \tag{11}$$

and Laplace-Stieltjes Transform (LST)

$$\Gamma^*(s) = \left(\frac{\mu}{\mu + s}\right)^{\alpha} \quad (Re(s) > 0). \tag{12}$$

Note that in the definition of the gamma-distribution μ is a scaling parameter, and that $\Gamma(\alpha, \mu)$ has the same distribution as $\mu^{-1}\Gamma(\alpha, 1)$. Using these definitions, the following result holds:

Theorem 1
Assume that all derivatives of $f(\underline{z})$ through order two exist at $\underline{z} = \underline{1}$ and that $0 < g_i < \infty$ $(i = 1, \ldots, M)$. Then

$$\lim_{n\to\infty} \frac{1}{\pi_n(\xi)} \begin{pmatrix} Z_n^{(1)} \\ \vdots \\ Z_n^{(M)} \end{pmatrix} \to_d A \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_M \end{pmatrix} \Gamma(\alpha, 1) \qquad (\xi \uparrow 1) \tag{13}$$

where $\underline{\hat{v}} = (\hat{v}_1, \ldots, \hat{v}_M)$ is the normalized the left eigenvector of $\hat{\mathbf{M}}$, and where $\Gamma(\alpha, 1)$ is a gamma-distributed random variable with scale parameter 1 and shape parameter

$$\alpha := \frac{1}{A} \underline{\hat{g}}^\top \underline{\hat{w}} = \frac{1}{A} \sum_{i=1}^{M} \hat{g}_i \hat{w}_i, \ \textit{with} \ \ A := \sum_{i=1}^{M} \hat{v}_i \left(\underline{\hat{w}}^\top \hat{\mathbf{K}}^{(i)} \underline{\hat{w}} \right) > 0. \tag{14}$$

Proof: See [26] (Theorem 4). □

In the next section we will show how this result, which was derived in the context of generic MTBPs, can be transformed into results for a general class of polling models.

3 Heavy-traffic Asymptotics for Polling Models

In this section we show how Theorem 1 can be transformed to derive new closed-form expressions for the LST of the queue-length and waiting-time distributions for a broad class of polling models, under heavy-traffic scalings. To this end, we consider two classical models that have been widely studied in the literature. In Section 3.1 we derive the LST of the asymptotic waiting-time distribution for cyclic polling models with globally-gated (GG) service. In Section 3.2 we derive asymptotic expressions for cyclic polling models with general branching-type service policies. In Section 3.3 we discuss the implications, the generality and the limitations of the results.

To avoid duplication, the following model assumptions and notation are introduced for both type of models. Consider an asymmetric cyclic polling model that consists of $N \geq 2$ queues, $Q_1, \ldots, Q_N$, and a single server that visits the queues in cyclic order. Customers arrive at Q_i accoring to a Poisson process with rate λ_i, and are referred to as type-i customers. The total arrival rate is $\Lambda := \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i, with LST $B_i^*(\cdot)$ and k-th moment $b_i^{(k)}$, which is assumed to be finite for $k = 1, 2$. The k-th moment of the service time of an arbitrary customer is $b^{(k)} := \sum_{i=1}^{N} \lambda_i b_i^{(k)} / \Lambda$ $(k = 1, 2)$. The total load of the system in $\rho := \sum_{i=1}^{N} \rho_i$. We define a polling instant at Q_i to be the moment at which

visit time at Q_i is defined as the time elapsed between a polling instant and its successive departure epoch at Q_i. Moreover, as i-cycle is the time between two successive polling instants at Q_i. Upon departing from Q_i the server immediately proceeds to Q_{i+1}, incurring a switch-over time R_i with LST $R_i^*(\cdot)$ and first two moments $r_i^{(k)}$ $(k = 1, 2)$, which are assumed to be finite. Denote by $r > 0$ and $r^{(2)} > 0$ be the first two moments of the switch-over time per 1-cycle of the server along the queues. The interarrivel times, service times and switch-over times are assumed to be mutually independent and independent of the state of the system.

Throughout, we focus on the behavior of the model when the load ρ tends to 1. For ease of the discussion we assume that as ρ changes the total arrival rate changes while the service-time distributions and ratios between the arrival rates are kept fixed; note that in this way, the limit for $\rho \uparrow 1$, which will be used frequently throughout this paper, is uniquely defined. Similar to the hat-notation for the MTBPs defined in Section 2, for each variable x that is a function of ρ we use the hat-notation $\hat{x}$ to indicate its value *at* $\rho = 1$.

For both models to be discussed below, joint queue-length vector at successive moments when the server arrives at a fixed queue (say Q_k) consitutes an MTBPs with immigration. To this end, the following notation is useful. Let $X_{i,n}^{(k)}$ be the number of type-i customers in the system at the n-th polling instant at Q_k, for $i, k = 1, \ldots, N$ and $n = 0, 1, \ldots$, and let $\underline{X}_n^{(k)} = (X_{1,n}^{(k)}, \ldots, X_{N,n}^{(k)})$ be the joint queue-length vector at the n-th pollling instant at Q_k. Moreover, $\mathbf{X}^{(k)} = \{\underline{X}_n^{(k)}, n = 0, 1, \ldots\}$ is the MTBP describing the evolution of the state of the system at successive polling instants at Q_k. For $\rho < 1$, we have $\underline{X}_n^{(k)} \to_d X^{(k)}$ for $n \to \infty$, where $X^{(k)}$ denotes the steady state joint queue-length vector at an arbitrary polling instant at Q_k.

3.1 Globally-Gated Service

The Globally-Gated (GG) service discipline works as follows (cf. [6]). At the beginning of a 1-cycle, marked by a polling instant at Q_1 (see above), all customers present at $Q_1, \ldots, Q_N$ are marked. During the coming 1-cycle (i.e., the visit of queues $Q_1, \ldots, Q_N$), the server serves all (and only) the marked customers. Customers that meanwhile arrive at the queues will have to wait until being marked at the next cycle-beginning, and will be served during the next 1-cycle. Since at each cycle the server serves all the work that arrived during the previous cycle, the stability condition is $\rho < 1$, which is both necessary and sufficient (cf. [13, 6]). Throughout this paper, this model will be referred to as the GG-model.

In 3.1.1 we show how Theorem 1 can be used to derive expressions for the LST of the asymptotic scaled waiting-time distributions at each of the queues. In 3.1.2 we discuss several interesting implications that follow from these expressions.

3.1.1 Analysis

To analyze the heavy-traffic behavior of the GG-model, we establish the relation with the general MTBP-model described in Section 2. To this end, recall that for the model considered here the joint queue-length process at embedded polling instants at Q_k (for any k) can be described as an N-dimensional MTBP with immigration in each state. For notational ease of the discussion that will follow, we proceed along two steps. First we focus on the heavy-traffic asymptotics for the joint queue-length vector at the successive moment at which the server arrives at Q_1 (Theorem 2). Second, we will transform these results to the joint queue-length distribution at polling instants at $Q_k,\ k = 1, \ldots, N$ (Theorem 3).

To start, we consider the MTBP $\mathbf{X}^{(1)} := \{\underline{X}_n^{(1)}, n = 0, 1, \ldots\}$ describing the evolution of the joint queue-legnth vector at successive polling instants of the server at Q_1. Then the process $\mathbf{X}^{(1)}$ is characterized by the offspring generating functions, for $i = 1, \ldots, N$,

$$f^{(i)}(z_1, \ldots, z_N) = B_i^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right) \tag{15}$$

and the immigration function

$$g(z_1, \ldots, z_N) = \prod_{i=1}^{N} R_i^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right). \tag{16}$$

$$g_j = \sum_{i=1}^{N} r_i \lambda_j = r\lambda_j. \tag{17}$$

To derive the limiting distribution of the joint queue-length vector at polling instants at Q_1, we need to specify the following parameters: (1) the mean matrix $\mathbf{M}$ and its corresponding left and right eigenvectors $\underline{\hat{v}}$ and $\underline{\hat{w}}$ at $\rho = 1$ (normalized according to (8)), and (2) the parameters A and $\underline{\hat{g}}$. These parameters are obtained in the following two Lemmas.

Lemma 1
For the GG-model, the mean matrix $\mathbf{M}$ *is given by the following expression:*

$$\mathbf{M} = \begin{pmatrix} b_1^{(1)}\lambda_1 & b_1^{(1)}\lambda_2 & \cdots & b_1^{(1)}\lambda_N \\ b_2^{(1)}\lambda_1 & \cdots & \cdots & b_2^{(1)}\lambda_N \\ \vdots & \vdots & \vdots & \vdots \\ b_N^{(1)}\lambda_1 & \cdots & \cdots & b_N^{(1)}\lambda_N \end{pmatrix} \text{ and hence, } \hat{\mathbf{M}} = \begin{pmatrix} b_1^{(1)}\hat{\lambda}_1 & b_1^{(1)}\hat{\lambda}_2 & \cdots & b_1^{(1)}\hat{\lambda}_N \\ b_2^{(1)}\hat{\lambda}_1 & \cdots & \cdots & b_2^{(1)}\hat{\lambda}_N \\ \vdots & \vdots & \vdots & \vdots \\ b_N^{(1)}\hat{\lambda}_1 & \cdots & \cdots & b_N^{(1)}\hat{\lambda}_N \end{pmatrix}. \tag{18}$$

Moreover, the right and left eigenvectors of $\hat{\mathbf{M}}$ *(i.e.,* $\mathbf{M}$ *at* $\rho = 1$*) are*

$$\underline{\hat{w}} = |\underline{b}|^{-1} \begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_N^{(1)} \end{pmatrix}, \quad \text{and} \quad \underline{\hat{v}} = |\underline{b}| \begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \\ \vdots \\ \hat{\lambda}_N \end{pmatrix}, \text{ respectively,} \tag{19}$$

with

$$\underline{b} := (b_1^{(1)}, \ldots, b_N^{(1)})^\top, \text{ and } |\underline{b}| := \sum_{i=1}^{N} b_i^{(1)}. \tag{20}$$

Proof: The first equation of (18) follows directly from (15) by differentiation: For $i, j = 1, \ldots, N$,

$$m_{i,j} := \frac{\partial f^{(i)}(\underline{z})}{\partial z_j}|_{\underline{z}=\underline{1}} = \frac{\partial}{\partial z_j} B_i^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right) |_{\underline{z}=\underline{1}} = b_i^{(1)} \lambda_j, \tag{21}$$

and the second equation in (18) then follows directly by evaluating the first equation at $\rho = 1$. To prove that $\underline{\hat{w}}$ is a right eigenvector of $\hat{\mathbf{M}}$, note that it follows directly from (18) that, for $i = 1, \ldots, N$,

$$\sum_{j=1}^{N} b_i^{(1)} \hat{\lambda}_j b_j^{(1)} = b_i^{(1)} \hat{\rho} = b_i^{(1)}, \tag{22}$$

so that $\hat{\mathbf{M}}\underline{b} = \underline{b}$, and hence, $\hat{\mathbf{M}}\underline{\hat{w}} = \underline{\hat{w}}$. Similarly, to show that $\underline{\hat{v}}$ is a left eigenvector of $\hat{\mathbf{M}}$, note that for $i = 1, \ldots, N$,

$$\sum_{j=1}^{N} \hat{\lambda}_j b_j^{(1)} \hat{\lambda}_i = \hat{\rho} \hat{\lambda}_i = \hat{\lambda}_i, \tag{23}$$

which implies $\hat{\mathbf{M}}^\top \underline{\hat{v}} = \underline{\hat{v}}$. This completes the proof of Lemma 1, by properly normalizing the eigenvectors according to (8). □

Lemma 2
For the GG-model, we have

$$\underline{\hat{g}}^\top \underline{\hat{w}} = |\underline{b}|^{-1} r, \tag{24}$$

and

$$A = |\underline{b}|^{-1} \frac{b^{(2)}}{b^{(1)}}. \tag{25}$$

$$\hat{\underline{g}}^\top \hat{\underline{w}} := \sum_{i=1}^{N} \hat{g}_i \hat{w}_i = \sum_{i=1}^{N} r|\underline{b}|^{-1} \hat{\lambda}_i b_i^{(1)} = \hat{\rho}|\underline{b}|^{-1} r = |\underline{b}|^{-1} r, \tag{26}$$

which follows directly from (17) and (19), and using the fact that $\hat{\rho} = 1$ by definition. To prove (25), we first observe that by differentiating (15) two times we have, for $i = 1, \ldots, N$,

$$\mathbf{K}^{(i)} = b_i^{(2)} \begin{pmatrix} \lambda_1^2 & \lambda_1\lambda_2 & \cdots & \lambda_1\lambda_N \\ \lambda_2\lambda_1 & \lambda_2^2 & \cdots & \lambda_2\lambda_N \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_N\lambda_1 & \cdots & \cdots & \lambda_N^2 \end{pmatrix}, \text{ and so } \hat{\mathbf{K}}^{(i)} = b_i^{(2)} \begin{pmatrix} \hat{\lambda}_1^2 & \hat{\lambda}_1\hat{\lambda}_2 & \cdots & \hat{\lambda}_1\hat{\lambda}_N \\ \hat{\lambda}_2\hat{\lambda}_1 & \hat{\lambda}_2^2 & \cdots & \hat{\lambda}_2\hat{\lambda}_N \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\lambda}_N\hat{\lambda}_1 & \cdots & \cdots & \hat{\lambda}_N^2 \end{pmatrix}. \tag{27}$$

Consequently, using (19) we have for $i = 1, \ldots, N$,

$$\hat{\underline{w}}^\top \hat{\mathbf{K}}^{(i)} \hat{\underline{w}} = |\underline{b}|^{-2} b_i^{(2)} \sum_{j=1}^{N} \sum_{k=1}^{N} b_j^{(1)} \hat{\lambda}_j \hat{\lambda}_k b_k^{(1)} = |\underline{b}|^{-2} b_i^{(2)}, \tag{28}$$

and hence, combining (19) and (28) we have

$$A := \sum_{i=1}^{N} \hat{v}_i \left(\hat{\underline{w}}^\top \hat{\mathbf{K}}^{(i)} \hat{\underline{w}} \right) = |\underline{b}|^{-1} \sum_{i=1}^{N} \hat{\lambda}_i b_i^{(2)} = |\underline{b}|^{-1} \hat{\Lambda} b^{(2)} = |\underline{b}|^{-1} \frac{b^{(2)}}{b^{(1)}}, \tag{29}$$

where the last equality follows form the fact that $\hat{\Lambda} = 1/b^{(1)}$. This completes the proof of Lemma 2. □

Let us consider the heavy-traffic behavior of the maximum eigenvalue ξ of $\mathbf{M}$. Note that in general, ξ is a non-negative real-valued function of ρ (cf. [2]), say

$$\xi = \xi(\rho), \tag{30}$$

for $\rho \geq 0$. Then the following result describes the behavior of $\xi(\cdot)$ in the neighbourhood of $\rho = 1$.

Lemma 3
For the GG-model, the maximum eigenvalue $\xi = \xi(\rho)$ has the following properties:

(1) $\xi < 1$ *if and only if* $0 \leq \rho < 1$, $\xi = 1$ *if and only if* $\rho = 1$, *and* $\xi > 1$ *if and only if* $\rho > 1$;
(2) $\xi = \xi(\rho)$ *is a continuous function of* ρ;
(3) $\lim_{\rho \uparrow 1} \xi(\rho) = f(1) = 1$;
(4) *the derivative of* $\xi(\cdot)$ *at* $\rho = 1$ *is given by*

$$\xi'(1) := \lim_{\rho \uparrow 1} \frac{1 - \xi(\rho)}{1 - \rho} = 1. \tag{31}$$

Proof: See Appendix A. □

We are now ready to transform Theorem 1 to the model under consideration.

Theorem 2
For the GG-model, the steady-state joint queue-length distribution at polling instants at Q_1 satisfies the following limiting behavior:

$$(1 - \rho) \begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \rightarrow_d \frac{b^{(2)}}{b^{(1)}} \begin{pmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_N \end{pmatrix} \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{32}$$

where

$$\alpha = r \frac{b^{(1)}}{b^{(2)}}. \tag{33}$$

Proof: First, it is readily verified that the joint-queue-length process $\mathbf{X}^{(1)} := \{\underline{X}_n^{(1)} = (X_{1,n}^{(1)}, \ldots, X_{N,n}^{(1)}), n = 0, 1, \ldots\}$ at embedded polling instants at Q_1 constitutes an N-dimensional MTBP with offspring function

Moreover, is it easy to verify that the assumptions of Theorem 1 are satisfied (with $M = N$). Then using Lemmas 1 to 3 and Theorem 1 it follows that

$$\lim_{n\to\infty} \frac{1}{\pi_n(\xi(\rho))} \begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} \to_d \frac{b^{(2)}}{b^{(1)}} \begin{pmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_M \end{pmatrix} \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{34}$$

where α is defined in (33). Consequently, relation (32) follows from the following sequence of equations:

$$\lim_{\rho\uparrow 1} (1-\rho) \begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} = \lim_{\rho\uparrow 1} \lim_{n\to\infty} (1-\rho) \begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} = \lim_{\rho\uparrow 1} \lim_{n\to\infty} (1-\rho)\pi_n(\xi(\rho)) \cdot \frac{1}{\pi_n(\xi(\rho))} \begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} \tag{35}$$

$$= \lim_{\rho\uparrow 1} \lim_{n\to\infty} (1-\rho)\pi_n(\xi(\rho)) \cdot \lim_{\rho\uparrow 1} \lim_{n\to\infty} \frac{1}{\pi_n(\xi(\rho))} \begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} = 1 \cdot \frac{b^{(2)}}{b^{(1)}} \begin{pmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_N \end{pmatrix} \Gamma(\alpha, 1), \tag{36}$$

where the last equality in (36) follows from Theorem 1 and the fact that (10) implies

$$\lim_{\rho\uparrow 1} \lim_{n\to\infty} (1-\rho)\pi_n(\xi(\rho)) = \lim_{\rho\uparrow 1} \frac{1-\rho}{1-\xi(\rho)} \cdot \lim_{n\to\infty} \frac{1-(\xi(\rho))^n}{\xi(\rho)} = 1 \cdot 1 = 1, \tag{37}$$

by using the properties formulated in Lemma 3. □

The next result generalizes Theorem 2, which gives the asymptotic scaled queue-length distribution at an arbitrary polling instant at Q_1, to the aymptotic queue-length distribution at an arbitrary polling instant at Q_k $(k = 1, \ldots, N)$.

Theorem 3
For the GG-model, the steady-state joint queue-length distribution at polling instants at Q_k $(k = 1, \ldots, N)$ satisfies the following limiting behavior:

$$(1-\rho) \begin{pmatrix} X_1^{(k)} \\ \vdots \\ X_N^{(k)} \end{pmatrix} \to_d \frac{b^{(2)}}{b^{(1)}} \left[(\hat{\rho}_1 + \cdots + \hat{\rho}_{k-1}) \begin{pmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_{k-1} \\ \hat{\lambda}_k \\ \vdots \\ \hat{\lambda}_N \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \hat{\lambda}_k \\ \vdots \\ \hat{\lambda}_N \end{pmatrix} \right] \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{38}$$

where

$$\alpha = r\frac{b^{(1)}}{b^{(2)}}. \tag{39}$$

Proof: For $k = 1, \ldots, N$, denote by $X_k^*(z_1, \ldots, z_N)$ the PGF of $(X_1^{(k)}, \ldots, X_N^{(k)})$, the joint queue length at an arbitrary polling instant at Q_k. Then it is readily verified that, for $|z_i| \leq 1$, $i = 1, \ldots, N$, $k = 1, \ldots, N$,

$$X_k^*(z_1, \ldots, z_N) = \prod_{i=1}^{k-1} R_i^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right) \tag{40}$$

$$\times\ X_1^* \left(B_1^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right), \ldots, B_{k-1}^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right), z_k, z_{k+1}, \ldots, z_N \right). \tag{41}$$

To this end, consider the customer population at a polling instant P_k^* at Q_k $(k > 1)$; note that for $k = 1$ the result was shown in Theorem 2 and is therefore not considered here again. Then this population consists of three independent parts: (1) the customers that were present at Q_i $(i = k, k+1, \ldots, N)$ at the last preceding polling instant at Q_1, (2) the customers who arrived during the service times of the customers that were present at Q_i $(i = 1, 2, \ldots, k-1)$ at the preceding polling instant at Q_1, and (3) the customers

standard generating function manipulations. Theorem 3 then follows directly from Theorem 2 by using (32) and taking the proper limits. □

We are now ready to obtain the main result for the GG-model.

Theorem 4
For the GG-model, the waiting-time distribution satisfies the following limiting behavior: For $i = 1, \ldots, N$,

$$(1-\rho)W_i \to_d \tilde{W}_i \quad (\rho \uparrow 1) \tag{42}$$

where the LST of $\tilde{W}_i$ is given by, for $Re(s) > 0$,

$$\tilde{W}_i^*(s) = \frac{1}{(1-\hat{\rho}_i)rs}\left\{\left(\frac{\mu}{\mu + s(\hat{\rho}_1 + \cdots + \hat{\rho}_i)}\right)^{\alpha} - \left(\frac{\mu}{\mu + s(1 + \hat{\rho}_1 + \cdots + \hat{\rho}_{i-1})}\right)^{\alpha}\right\}, \tag{43}$$

where

$$\alpha = r\frac{b^{(1)}}{b^{(2)}}, \text{ and } \mu = \frac{b^{(1)}}{b^{(2)}}. \tag{44}$$

Proof: Denote by $X_i^{(i)}$ and $Y_i^{(i)}$ the number of customers at Q_i at the beginning and at the end of a visit period to Q_i, respectively, and denote by N_i be the number of customers at Q_i at an arbitrary customer departure epoch from Q_i. Denote the corresponding PGFs by $X_i^*(\cdot)$, $Y_i^*(\cdot)$ and $N_i^*(\cdot)$. Then the following result was obtained by Borst and Boxma [8]: For $|z| \leq 1$, $i = 1, \ldots, N$,

$$N_i^*(z) = \frac{(1-\rho_i)(1-z)B_i^*(\lambda_i(1-z))}{B_i^*(\lambda_i(1-z)) - z}\frac{Y_i^*(z) - X_i^*(z)}{(1-z)\lambda_i(1-\rho_i)r/(1-\rho).} \tag{45}$$

Then from Theorem 3, taking the i-th component only, we have that in the limiting case $\rho \uparrow 1$,

$$(1-\rho)X_i^{(i)} \to_d \frac{b^{(2)}}{b^{(1)}} \cdot \hat{\lambda}_i(1 + \hat{\rho}_1 + \cdots + \hat{\rho}_{i-1}) \cdot \Gamma(\alpha, 1). \tag{46}$$

Then, to determine the number of type-i customers $Y_i^{(i)}$ at the end of a visit of the server to Q_i, note that during the visit of the server to Q_i, each of the $X_i^{(i)}$ customers is effectively replaced a joint set of customers with PGF $B_i^*\left(\sum_{j=1}^N \lambda_j(1-z_j)\right)$; focusing on customers at Q_i only, it is readily seen that at the end of a visit period to Q_i each of the type-i customers present at the beginning of that visit period has been effectively by a number of type-i customers with marginal PGF $B_i^*(\lambda_i(1-z_i))$, with average ρ_i, which is easily seen to imply that in the limiting case $\rho \uparrow 1$,

$$(1-\rho)Y_i^{(i)} \to_d \frac{b^{(2)}}{b^{(1)}} \cdot \hat{\lambda}_i(\hat{\rho}_1 + \cdots + \hat{\rho}_i) \cdot \Gamma(\alpha, 1). \tag{47}$$

Combining (45)-(47), using the distributional form of Little's formula and the observation that a departing customer sees the time average [32] is then easily seen to lead to (43), which completes the proof of Theorem 4. □

3.1.2 Implications

Theorem 4 leads to a number of interesting implications that will be discussed below.

Corrollary 1 (Insensitivity properties)
For $i = 1, \ldots, N$, the asymptotic waiting-time distribution $\tilde{W}_i$,

(1) *is independent of the visit order (assuming the order is cyclic),*
(2) *depends on the variability of the service-time distributions only through $b^{(2)}$, and*
(3) *depends on the switch-over time distributions only through r.*

Note that similar insensitivity properties are generally not valid for stable systems (i.e., $\rho < 1$), in which case the waiting-time distributions *do* depend on the visit order, the complete service-time distributions and each of the individual switch-over time distributions. Apparently, these dependencies are of lower order, and hence their effect on the waiting-time distributions becomes negligible, in heavy traffic.

For the case of zero switch-over times, the LST of $\tilde{W}_i$ for the GG-model is given by the following expression: For $i = 1, \ldots, N$, $Re(s) \geq 0$,

$$\lim_{r \downarrow 0} \tilde{W}_i(s) = \frac{1}{(1-\hat{\rho}_i)s} \frac{b^{(1)}}{b^{(2)}} \log\left(\frac{\mu + s(1 + \hat{\rho}_1 + \cdots + \hat{\rho}_{i-1})}{\mu + s(\hat{\rho}_1 + \cdots + \hat{\rho}_i)} \right), \tag{48}$$

where α and μ are defined in (44), and where $\log(\cdot)$ is an inverse function of the (complex) function $l(z) := \exp(z)$.

Corrollary 3 (Expected asymptotic delay)
For the GG-model, the asymptotic expected delay at Q_i is given by the following expression: For $i = 1, \ldots, N$,

$$E[\tilde{W}_i] = \frac{1}{2} \left(1 + 2 \sum_{j=1}^{i-1} \hat{\rho}_j + \hat{\rho}_i \right) \left(\frac{b^{(2)}}{b^{(1)}} + r \right). \tag{49}$$

Remark 1 (Pseudo-conservation law):
The pseudo-conservation law (PCL) for the present model is as follows (cf. [6]): For $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i E[W_i] = \rho \sum_{i=1}^{N} \frac{\lambda_i b_i^{(2)}}{2(1-\rho)} + \rho \frac{r^{(2)}}{2r} + \rho^2 \frac{r}{1-\rho} + \sum_{i=2}^{N} \rho_i \sum_{j=1}^{i-1} r_j^{(1)}. \tag{50}$$

By taking heavy-traffic limits, it follows directly that

$$\sum_{i=1}^{N} \rho_i E[\tilde{W}_i] = \frac{b^{(2)}}{2b^{(1)}} + r. \tag{51}$$

Then it is easy to verify that equation (49) indeed satisfies (51), which supports the validity of Theorem 4.

3.2 Cyclic polling models with general branching-type service policies

In this section we consider the cyclic polling model introduced at the beginning of Section 3, with general branching-type service policies that satisfy the following property (cf. [29]):

Branching property
If the server arrives at Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a polulation of customers having joint probability generating function (PGF) $h_i(\underline{z}) = h_i(z_1, \ldots, z_N)$, which can be any N-dimensional PGF.

We assume that the service disciplines are *work conserving*, in the sense that the server always works during a visit to a queue. From the branching property, a visit period of the server starting with k_i original customers, say $C_1, \ldots, C_{k_i}$, consist of k_i mutually independent sub-busy period, each of which is characterized by the joint PGF-LST: For $i = 1, \ldots, N$, $Re(u) > 0$, $|v| \leq 1$,

$$\psi_i(u, v) := E\left[e^{-uT_i} v^{L_i} \right], \tag{52}$$

where T_i is the duration of a sub-busy period, and L_i is the so-called sub-busy period residue, i.e., the number of type-i children of the original customer that generates this sub-busy period.

This class of service policies contains a variety of classical service policies, including the exhaustive, gated, binomial-gated [19] and binomial-exhaustive [29] policies, amongst others. For gated and exhaustive service at Q_i, we have for $|z_k| \leq 1$ $(k = 1, \ldots, N)$,

$$h_i(\underline{z}) = B_i^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right) \quad \text{and} \quad h_i(\underline{z}) = \Theta_i^* \left(\sum_{j \neq i} \lambda_j (1 - z_j) \right), \tag{53}$$

respectively, where $\Theta_i^*(\cdot)$ denotes the LST of a busy period in an $M/G/1$ queue with arrival rate λ_i and service time distribution $B_i^*(\cdot)$. Similarly, for the case of binomial-gated service (with parameter $0 < p_i \leq 1$) and binomial-exhaustive service (with parameter $0 < q_i \leq 1$) we have for $|z_k| \leq 1$ $(k = 1, \ldots, N)$,

$$h_i(\underline{z}) = (1 - p_i) z_i + p_i B_i^* \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right) \quad \text{and} \quad h_i(\underline{z}) = (1 - q_i) z_i + q_i \Theta_i^* \left(\sum_{j \neq i} \lambda_j (1 - z_j) \right), \tag{54}$$

at Q_i $(i = 1, \ldots, N)$ by

$$f_i := 1 - E[L_i], \tag{55}$$

where L_i is the sub-busy period residue, defined in (52). The exhaustiveness f_i has the following simple interpretation: each customer present at Q_i at the beginning of a visit of the service Q_i is effectively replaced by a number of customers at Q_i whose mean value is $1 - f_i$. In other words, f_i can be seen as

$$1 - f_i := \frac{\text{E[number of customers at } Q_i \text{ at the end of a visit to } Q_i]}{\text{E[number of customers at } Q_i \text{ at the beginning of that visit to } Q_i]}. \tag{56}$$

It is readily verified from equations (52)-(56) that for the case of exhaustive and gated service we have $f_i = 1$ and $f_i = 1 - \rho_i$, respectively (see also Remark 2 below). Notice also that the work conserving property implies the following relation between the sub-busy period duration T_i and the sub-busy period residue L_i: For $i = 1, \ldots, N$,

$$E[T_i] = (1 - E[L_i])\frac{b_i^{(1)}}{1 - \rho_i} = f_i \frac{b_i^{(1)}}{1 - \rho_i}. \tag{57}$$

3.2.1 Analysis

To establish the relation with the general MTBP-model described in Section 2, we observe that for the model considered here the joint queue-length process at embedded polling instants at Q_1 can be described as an N-dimensional MTBP with immigration in each state (cf. [29]). This process is characterized by the offspring generating functions, for $|z_k| \leq 1$ $(k = 1, \ldots, N)$, $i = 1, \ldots, N$,

$$f^{(i)}(z_1, \ldots, z_N) = h_i(z_1, z_2, \ldots, z_i, f^{(i+1)}(z_1, \ldots, z_N), \ldots, f^{(N)}(z_1, \ldots, z_N)), \tag{58}$$

with

$$h_i(z_1, \ldots, z_N) := \psi_i \left(\sum_{j \neq i} \lambda_j (1 - z_j), z_i \right), \tag{59}$$

where $\psi_i(\cdot, \cdot)$ is defined in (52), and the immigration function, for $|z_k| \leq 1$ $(k = 1, \ldots, N)$,

$$g(z_1, \ldots, z_N) = \prod_{i=1}^{N} R_i^* \left(\sum_{k=1}^{i} \lambda_k (1 - z_k) + \sum_{k=i+1}^{N} \lambda_k (1 - f^{(k)}(z_1, \ldots, z_N)) \right). \tag{60}$$

To derive the limiting distribution of the joint queue-length vector at polling instants at Q_1, we need to specify the following parameters: (1) the mean matrix $\hat{\mathbf{M}}$ and its corresponding (normalized) left and right eigenvectors $\underline{\hat{v}}$ and $\underline{\hat{w}}$, and (2) the parameters $\hat{g}$ and A. These parameters are obtained in the following two lemmas.

Lemma 4
For the cyclic branching-type polling model, the mean matrix $\mathbf{M}$ *is given by the following expression:*

$$\mathbf{M} = \mathbf{M_1} \cdots \mathbf{M}_N, \tag{61}$$

where for $i = 1, \ldots, N$,

$$\mathbf{M}_i = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & \vdots & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_1 f_i \varphi_i & \lambda_2 f_i \varphi_i & \cdots & \lambda_{i-1} f_i \varphi_i & 1 - f_i & \lambda_{i+1} f_i \varphi_i & \vdots & \vdots & \lambda_N f_i \varphi_i \\ 0 & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & 1 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{62}$$

$$u_i := \frac{\lambda_i(1-\rho_i)(1-f_i)}{f_i} + \lambda_i \sum_{j=i+1}^{N} \rho_j, \tag{63}$$

then the normalized right and left eigenvectors of $\hat{\mathbf{M}}$ *are given by*

$$\underline{\hat{w}} = \begin{pmatrix} \hat{w}_1 \\ \vdots \\ \hat{w}_N \end{pmatrix} = |\underline{b}|^{-1} \begin{pmatrix} b_1^{(1)} \\ \vdots \\ b_N^{(1)} \end{pmatrix}, \quad \textit{and } \underline{\hat{v}} = \frac{|\underline{b}|}{\delta} \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_N \end{pmatrix}, \tag{64}$$

with

$$\delta := \underline{\hat{u}}^\top \underline{\hat{w}} = \sum_{i=1}^{N} \left(\frac{\hat{\rho}_i(1-\hat{\rho}_i)(1-\hat{f}_i)}{\hat{f}_i} + \hat{\rho}_i \sum_{j=i+1}^{N} \hat{\rho}_j \right), \tag{65}$$

and where $\underline{b}$ *and* $|\underline{b}|$ *are defined in* (20)*.*

Proof: To prove (61) and (62), consider a tagged type-i customer, say C_i, present at Q_i at the beginning of a service period at Q_i. Following the branching property, C_i generates a sub-busy period with joint PGF-LST $\psi_i(\cdot,\cdot)$, defined in (52). During this sub-busy period, the average number of children C_i has at Q_j ($j \neq i$) is $\lambda_j E[T_i] = \lambda_j f_i \varphi_i$, by using (57). Moreover, it is readily seen that the number of type-i children of C_i is exactly the residue of the sub-busy period generated by C_i, and its mean value equals $E[L_i] = 1 - f_i$. Based on these observations, equations (61) and (62) are easily seen to hold, for $i = 1, \ldots, N$. To proof that $\underline{\hat{w}}$ is a right eigenvector at $\hat{\mathbf{M}}$, note that it follows directly from (62) that, for $i = 1, \ldots, N$,

$$\sum_{j \neq i} \hat{\lambda}_j \hat{f}_i \hat{\varphi}_i b_j^{(1)} + b_i^{(1)}(1-\hat{f}_i)\hat{f}_i\hat{\varphi}_i \sum_{j \neq i} \hat{\lambda}_j b_j^{(1)} + b_i^{(1)}(1-\hat{f}_i)\hat{f}_i b_i^{(1)} + b_i^{(1)}(1-\hat{f}_i) = b_i^{(1)}, \tag{66}$$

so that $\hat{\mathbf{M}}_i \underline{\hat{w}} = \underline{\hat{w}}$ ($i = 1, \ldots, N$), and hence, $\hat{\mathbf{M}}\underline{\hat{w}} = \underline{\hat{w}}$, which shows that $\underline{\hat{w}}$ is indeed a right eigenvector of $\hat{\mathbf{M}}$. Similar arguments can be used to show that $\underline{\hat{v}}$ is a left eigenvector of $\hat{\mathbf{M}}$ (along the lines discussed in the Appendix of [37]). The details are omitted for compactness of the presentation, and are left as an exercise to the reader. This completes the proof of Lemma 4. □

Lemma 5
For the cyclic branching-type polling model,

$$\underline{\hat{g}}^\top \underline{\hat{w}} = |\underline{b}|^{-1} r, \tag{67}$$

and

$$A = |\underline{b}|^{-1} \delta^{-1} \cdot \frac{b^{(2)}}{b^{(1)}}. \tag{68}$$

Proof: Assume $\rho = 1$. To show (67) we first observe that it follows from (60) that the mean number of type-j customers that immigrate during a cycle is given by

$$\hat{g}_j = \sum_{i=1}^{N} r_i^{(1)} \left(\hat{\lambda}_j I_{\{j \leq i\}} + \sum_{k=i+1}^{N} \hat{\lambda}_k \hat{m}_{k,j} \right), \tag{69}$$

where I_E stands for the indicator function on the event E. This implies

$$\hat{g}^\top \hat{w} := \sum_{j=1}^{N} \hat{g}_j \hat{w}_j = |\underline{b}|^{-1} \sum_{j=1}^{N} \hat{g}_j b_j^{(1)} = |\underline{b}|^{-1} \sum_{i=1}^{N} r_i^{(1)} \left(\hat{\lambda}_j b_j^{(1)} I_{\{j \leq i\}} + \sum_{k=i+1}^{N} \hat{\lambda}_k \sum_{j=1}^{N} \hat{m}_{k,j} b_j^{(1)} \right) \tag{70}$$

$$= |\underline{b}|^{-1} r \sum_{i=1}^{N} \hat{\rho}_i = |\underline{b}|^{-1} r, \tag{71}$$

by using (64), (69), and the fact that $\sum_{j=1}^{N} \hat{m}_{k,j} b_j^{(1)} = b_k^{(1)}$, which is an immediate consequence of the second part of Lemma 1, see (19). Finally, the proof of (68) can be obtained along similar lines as the proof of (25) in (26)-(29), but with notationally cumbersome derivations. The details are omitted for compactness of the

Lemma 6
For the cyclic branching model, the maximum eigenvalue $\xi = \xi(\rho)$ has the following properties:

(1) $\xi < 1$ *if and only if* $\rho < 1$, $\xi = 1$ *if and only if* $\rho = 1$ *and* $\xi > 1$ *if and only if* $\rho > 1$;
(2) $\xi(\rho)$ *is a continuous function of* ρ;
(3) $\lim_{\rho\uparrow 1} \xi(\rho) = \xi(1) = 1$;
(4) *the derivative of* $\xi(\rho)$ *at* $\rho = 1$ *is given by*

$$\xi'(1) = \lim_{\rho\uparrow 1} \frac{1-\xi(\rho)}{1-\rho} = \frac{1}{\delta}, \tag{72}$$

where δ *is defined in* (65).

Proof: See Appendix B. □

We are now ready to present the main result for the model under consideration.

Theorem 5
For the cyclic branching-type polling model, the joint queue-length vector at polling instants at Q_1 *has the following asymptotic behavior:*

$$(1-\rho)\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \rightarrow_d \frac{b^{(2)}}{b^{(1)}}\ \frac{1}{\delta}\begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_N \end{pmatrix}\ \Gamma(\alpha,1) \quad (\rho\uparrow 1), \tag{73}$$

where

$$\alpha = r\delta\frac{b^{(1)}}{b^{(2)}}. \tag{74}$$

and where δ *and* $\hat{u}_i$ $(i = 1,\ldots,N)$ *are defined in* (65) *and* (63)*, respectively.*

Proof: To start, note that the joint-queue-length process $\mathbf{X}^{(1)} := \{\underline{X}_n^{(1)} = (X_{1,n}^{(1)},\ldots,X_{N,n}^{(1)}), n = 0,1,\ldots\}$ at embedded polling instants at Q_1 constitutes an N-dimensional MTBP with offspring function $f^{(i)}(\underline{z})$ and immigration function $g(\underline{z})$ defined in (58) and (60), and with mean matrix $\mathbf{M}$ defined in (61)-(62). Moreover, is it easy to verify that the assumptions of Theorem 1 are satisfied (with $M = N$). Then using Lemmas 4 to 6 and Theorem 1 it follows that

$$\lim_{n\to\infty}\ \frac{1}{\pi_n(\xi(\rho))}\begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} \rightarrow_d A\begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_N \end{pmatrix}\Gamma(\alpha,1) \quad (\rho\uparrow 1), \tag{75}$$

where α, $\underline{\hat{v}}$ and A are given in (74), (64) and (68), respectively. Hence, similar to the derivation of Theorem 2, relation (73) follows from the following sequence of equations:

$$\lim_{\rho\uparrow 1}\ (1-\rho)\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} = \lim_{\rho\uparrow 1}\lim_{n\to\infty}\ (1-\rho)\begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} \tag{76}$$

$$= \lim_{\rho\uparrow 1}\ \lim_{n\to\infty}\ (1-\rho)\pi_n(\xi(\rho)) \cdot \lim_{\rho\uparrow 1}\ \lim_{n\to\infty}\ \frac{1}{\pi_n(\xi(\rho))}\begin{pmatrix} X_{n,1}^{(1)} \\ \vdots \\ X_{n,N}^{(1)} \end{pmatrix} = \delta\cdot A\begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_N \end{pmatrix}\Gamma(\alpha,1) \tag{77}$$

$$= \frac{1}{\delta}\cdot\frac{b^{(2)}}{b^{(1)}}\begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_N \end{pmatrix}\Gamma(\alpha,1). \tag{78}$$

$$\lim_{\rho\uparrow 1}\lim_{n\to\infty}\ (1-\rho)\pi_n(\xi(\rho)) = \lim_{\rho\uparrow 1}\ \frac{1-\rho}{1-\xi(\rho)}\ \lim_{n\to\infty}\frac{1-(\xi(\rho))^n}{\xi(\rho)} = \delta\cdot 1 = \delta, \tag{79}$$

which follows directly by using (10) and the properties listed in Lemma 6. Finally, equation (78) follows from (64) and (68). This completes the proof of Theorem 5. □

Theorem 6
For the cyclic branching-type polling model, the waiting-time distribution satisfies the following limiting behavior: For $i = 1, \ldots, N$,

$$(1-\rho)W_i \to_d \tilde{W}_i \quad (\rho\uparrow 1) \tag{80}$$

where the LST of $\tilde{W}_i$ is given by

$$\tilde{W}_i^*(s) = \frac{1}{(1-\hat{\rho}_i)rs}\left\{\left(\frac{\mu_i}{\mu_i + s(1-\hat{f}_i)}\right)^{\alpha} - \left(\frac{\mu_i}{\mu_i+s}\right)^{\alpha}\right\} \quad (Re(s)>0), \tag{81}$$

where

$$\alpha = r\delta\frac{b^{(1)}}{b^{(2)}}, \quad \mu_i = \delta\frac{b^{(1)}}{b^{(2)}}\frac{\hat{f}_i}{1-\hat{\rho}_i}, \tag{82}$$

and where δ is given in (65).

Proof: Without loss of generality, we focus on the waiting time distribution at Q_1. Adopting the notation used in the proof of Theorem 4, relation (45) is also applicable to the cyclic branching-type model under consideration (and hence also for the special case $i = 1$), so it remains to determine the limiting behavior for $X_1^{(1)}$ and $Y_1^{(1)}$, i.e. the number of type-1 customers present at the beginning and the end of a visit period to Q_1, respectively. To this end, note that Theorem 4 implies that in the limiting case $\rho\uparrow 1$,

$$(1-\rho)X_1^{(1)} \to_d \frac{b^{(2)}}{b^{(1)}}\cdot\frac{1}{\delta}\cdot u_1\cdot\Gamma(\alpha,1). \tag{83}$$

Then, using the branching structure of the service policity at Q_1 it is then readily seen that, for $\rho\uparrow 1$,

$$(1-\rho)Y_1^{(1)} \to_d (1-\hat{f}_1)\cdot\frac{b^{(2)}}{b^{(1)}}\cdot\frac{1}{\delta}\cdot u_1\cdot\Gamma(\alpha,1). \tag{84}$$

To see the latter, note that at the end of the visit period V_1 at Q_1, each type-1 customer that was present at the beginning of V_1 has been replaced by a population of customers whose PGF is given by $\psi_1(\cdot,\cdot)$, defined in (52). Focusing on type-1 customers only, each type-1 customer present Q_1 at the beginning of V_1 is replaced by, on average, $1-f_1$ type-1 customers at the end of V_1. Then, combining (83)-(84), using the distributional form of Little's formula and the observation that a departing customer sees the time average [32] is easily seen to lead to (80)-(81), recalling that we assumed $i = 1$ without loss of generality. □

The results presented in Theorem 6 are new and have not been observed before in the general context of the model considered. We emphasize that the results are valid in the general parameter setting of the model defined above. Remarkably, the results can be obtained in closed form, and moreover, are strikingly simple, and explicitly show the impact of the system parameters on the asymptotic delay at each of the queues.

3.2.2 Implications

Theorem 6 leads to a number of interesting implications that will be addressed below.

Corollary 5 (Insensitivity properties)
For $i = 1, \ldots, N$, the asymptotic waiting-time distribution $\tilde{W}_i$,

(1) *depend on the service policies only through the exhaustiveness factors $f_1, \ldots, f_N$,*
(2) *is independent of the visit order (assuming the order is cyclic),*
(3) *depends on the variability of the service-time distributions only through $b^{(2)}$, and*
(4) *depends on the switch-over time distributions only through r.*

which case the waiting-time distributions depend on the complete distribution of the sub-busy periods defined in (52), the visit order, the complete service-time distributions and each of the individual switch-over time distributions. Apparently, these dependencies are of lower order, and hence their effect on the waiting-time distributions becomes negligible, in heavy traffic.

Corollary 6 (Zero switch-over times)
For the special case of zero switch-over times, we have: For $i = 1, \ldots, N$, $Re(s) > 0$,

$$\lim_{r \downarrow 0} \tilde{W}_i(s) = \frac{\delta}{(1-\hat{\rho}_i)s} \frac{b^{(1)}}{b^{(2)}} \log\left(\frac{\mu_i + s}{\mu_i + s(1-\hat{f}_i)} \right), \tag{85}$$

where α, μ_i *and* δ *are defined in* (82) *and* (65)*, respectively, and where* $\log(\cdot)$ *is an inverse function of the (complex) function* $l(z) := \exp(z)$.

Corollary 7 (Mean asymptotic delay)
For the cyclic branching model, the asymptotic expected delay at Q_i *is given by the following expression: For* $i = 1, \ldots, N$,

$$E[\tilde{W}_i] = \frac{(1-\hat{\rho}_i)\left(\frac{2}{\hat{f}_i} - 1\right)}{\sum_{j=1}^{N} \hat{\rho}_j (1-\hat{\rho}_j)\left(\frac{2}{\hat{f}_j} - 1\right)} \frac{b^{(2)}}{b^{(1)}} + \frac{1}{2} r (1-\hat{\rho}_i)\left(\frac{2}{\hat{f}_i} - 1\right). \tag{86}$$

Note that this result was also shown in [39], where we obtained the result via the Descendant Set Approach [16].

We end this subsection with a number of remarks.

Remark 2 (Generalization of known results):
Theorem 5 generalizes and unifies known results that have been shown before. Van der Mei [36] derived the result for the special case of mixtures of gated and exhaustive service at each queue: if E denotes the set of queues that receive exhaustive service and its complement G the denoted the set of queues that received gated service, then is readily verified from equation (53) that $f_i = 1$ for $i \in E$, and $f_i = 1 - \rho_i$ for $i \in G$, which is easily seen that in that case $\delta = \left(1 - \sum_{i \in E} \hat{\rho}_i^2 + \sum_{i \in G} \hat{\rho}_i^2\right)/2$. Similarly, from equation (54) it follows that for the case of binomial-gated service at Q_i with probability p_i $(0 < p_i \leq 1)$ occurs as a special case with $f_i = p_i(1-\rho_i)$, and for the fractional exhaustive policy with parameter q_i $(0 < q_i \leq 1)$ we have $f_i = q_i$; to the best of the author's knowledge these results have not been shown before in the literature.

Remark 3 (Pseudo-conservation law):
The pseudo-conservation law (PCL) for the present model is as follows (cf. [42]): For $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i E[W_i] = \rho \sum_{i=1}^{N} \frac{\lambda_i b_i^{(2)}}{2(1-\rho)} + \rho \frac{r^{(2)}}{2r} + \frac{r}{2(1-\rho)} \left[\rho^2 - \sum_{i=1}^{N} \rho_i^2 \right] + \sum_{i=1}^{N} E[M_i], \tag{87}$$

where the mean amount of at Q_i at a server departure instant at Q_i is, for $\rho < 1$, $i = 1, \ldots, N$,

$$E[M_i] = \frac{r\rho_i(1-\rho_i)(1-f_i)}{f_i(1-\rho)}. \tag{88}$$

By taking heavy-traffic limits, it follows directly that

$$\sum_{i=1}^{N} \rho_i E[\tilde{W}_i] = \frac{b^{(2)}}{2b^{(1)}} + \frac{r}{2} \sum_{i=1}^{N} \hat{\rho}_i (1-\hat{\rho}_i)(1-\hat{f}_i)\left(\frac{2}{\hat{f}_i} - 1\right). \tag{89}$$

Then it is easy to verify that equation (86) indeed satisfies (89), which supports the validity of Theorem 6.

Remark 4 (Direct calculation of mean values):
The mean values of $X_i^{(k)}$ $(i, k = 1, \ldots, N)$ can also be obtained directly via simple balancing arguments. To this end, note first that for $i = k$ simple balancing arguments lead to the following equations: For $\rho < 1$, $i = 1, \ldots, N$,

$$E[X_i^{(i)}] = \lambda_i r + \lambda_i \sum_{j \neq i} E[X_j^{(j)}] E[T_j] + E[X_i^{(i)}] E[L_i], \tag{90}$$

$$E[X_i^{(i)}] = \frac{r}{1-\rho}\frac{\rho_i}{E[T_i]} = \frac{\lambda_i r(1-\rho_i)}{f_i(1-\rho)}. \tag{91}$$

Notice that for the special case $i = 1$ it follows from Theorem 5 that

$$\lim_{\rho\uparrow 1}\ (1-\rho)E[X_1^{(1)}] = \frac{b^{(2)}}{b^{(1)}}\cdot\frac{1}{\delta}\cdot\hat{u}_1\cdot\alpha = \frac{b^{(2)}}{b^{(1)}}\cdot\frac{\hat{\lambda}_1(1-\hat{\rho}_1)}{\hat{f}_1}\cdot r\delta\frac{b^{(1)}}{b^{(2)}} = \frac{r\hat{\lambda}_1(1-\hat{\rho}_1)}{\hat{f}_1(1-\rho)}, \tag{92}$$

where the second equality follows from the fact that

$$\hat{u}_1 = \frac{\hat{\lambda}_1(1-\hat{\rho}_1)(1-\hat{f}_1)}{\hat{f}_1} + \hat{\lambda}_1(1-\hat{\rho}_1) = \frac{\hat{\lambda}_1(1-\hat{\rho}_1)}{\hat{f}_1}. \tag{93}$$

Note that equation (92) is indeed in line with (91). More generally, from simple balancing arguments it follows directly that, for $\rho < 1$, $i, k = 1, \ldots, N$,

$$E[X_i^{(k)}] = \frac{\lambda_i r(1-\rho_i)(1-f_i)}{f_i(1-\rho)} + \frac{\lambda_i r}{1-\rho}\sum_{j=i+1}^{k-1}\rho_j. \tag{94}$$

Then it is readily verified from Theorem 5 that for the case $k = 1$ (without loss of generality), for $i = 1, \ldots, N$,

$$\lim_{\rho\uparrow 1}\ (1-\rho)E[X_i^{(1)}] = \frac{1}{\delta}\cdot\frac{b^{(2)}}{b^{(1)}}\cdot\hat{u}_i\cdot\alpha = r\hat{u}_i = \frac{\hat{\lambda}_i r(1-\hat{\rho}_i)(1-\hat{f}_i)}{\hat{f}_i} + \hat{\lambda}_i r\sum_{j=i+1}^{N}\hat{\rho}_j, \tag{95}$$

which is in line with Theorem 5.

3.3 Discussion and further remarks

Model extensions: The results presented in Sections 3.1 and 3.2 can be readily extended to a broader set of models. The requirements for the derivation of heavy-traffic limits similar to Theorems 2 to 6 are that (1) the evolution of the system at specific moments can be described as a multi-dimensional branching process with immigration, and (2) that the systen is work conserving. In addition to the models addressed above, this class of models includes as special cases for example models with gated/exhaustive service and non-cyclic periodic server routing [43], models with (simultaneous) batch arrivals [39, 21], continuous polling models [17], models with customer routing [31], globally-gated models with elevator-type routing [1], models with local priorities [30], amongst many other model variants. Basically, all that needs to be done for each of these model variants is to determine the parameters α, $\underline{\hat{u}}$ and the derivative of $\xi = \xi(\rho)$ at $\rho = 1$, which is usually straightforward.

Generality of the results: The question raises which polling models fall within the class of branching-type models for which the approach presented is this paper is applicable. As stated above, the key requirements are the existence of a suitable embedded process such that the evolution of the state of the system can be described by an MTBP, and that the system is work conserving. Although most polling systems that are used in practice are indeed work conserving, it is not inconceivable that there exist non work-conserving polling models for which an embedded process does satisfy an MTBP-structure. In those cases, properties similar to those stated in Lemma's 4 and 6 are no longer valid, so that the translation of Theorem 1 to results for polling models similar to Theorems 3 and 5, which explicitly use Lemma's 4 and 6, may be more complicated. Moreover, the required MTBP-structure of a proper embedded processes implies that the arrival processes should be memoriless, and hence must be Poisson, or some batched variant of the Poisson process. For example, models with renewal processes with non-exponential interarrival times generally violate the required branching structure, and hence, fall beyond the scope of the branching-type models for which our results hold (see also the remarks about this in Section 4 below).

Choice of the embedded process: In general, the MTBP need not always be the joint queue-length vector at embedded polling instants at a fixed queue, with $M = N$. For example, in the case of periodic server routing with polling table $\underline{\pi} := (\pi_1, \ldots, \pi_L)$ of length $L \geq N$ a proper choice for the MTBP is the $M := L$-dimensional joint queue-length is a fixed *pseudo*-queue [43]. As another example, in the case of two-stage polling models with cyclic routing [25], one should most likely consider the $M := 2N$-dimensional state vector describing the numbers of customers at both stages of all N types at embedded polling instant at a fixed queue; here, the state of the system cannot be described completely by an N-dimensional state vector.

Assumptions on the finiteness of moments: Theorems 4 and 5 are valid under the assumption that the second moments of the service times and the first moments of the switch-over times are finite; these assumptions are an immediate consequence of the assumptions on the finiteness of the mean immigration function g and the second-order derivatives of the offspring function $K_{j,k}^{(i)}$, defined in (5) and (7), respectively. It is interesting to observe that the results obtained in by Van der Mei [36] via the use of the Descendant Set Approach (DSA) assumes the finiteness of *all* moments of the service times and switch-over times; these assumptions were required, since the DSA-based proofs in [36] are based on a *bottom-up* approach in the sense that the limiting results for the waiting-time distributions are obtained from the asymptotic expressions for the moments of the waiting times obtained in [38, 37]. Note that in this way the DSA-based approach differs fundamentally from the *top-down* approach taken in the present paper, where the asymptotic expressions for the moments can be obtained from the expressions for the asymptotic waiting-time distributions in Theorems 4 and 5.

Local and global branching: Although the GG-model discussed in Section 3.1 the joint queue-length vector at successive polling instants at a fixed queue constitutes an MTBP, the GG-model does not occur as a special case of the branching model discussed in Section 3.2. To this end, note that for the GG-model the service policy at Q_i does not satisfy the *local* branching property described in Section 3.2, for $i > 1$. To see this, consider an arbitrary polling instant at Q_i ($i > 1$), which marks the beginning of a visit V_i to Q_i. Then the number of customers present at that moment, say $L_i^{(total)}$, can be written as

$$L_i^{(total)} = L_i^{(front)} + L_i^{(behind)}, \tag{96}$$

where $L_i^{(front)}$, $L_i^{(behind)}$ stands for the number of type-i customers that *in front of* and *behind* the global gate, respectively. Then at the end of V_i all $L_i^{(front)}$ customers that were standing in front of the gate have been served and hence have been effectively replaced by a population of customers whose joint PGF is given by $B_i^*(\sum_{j=1}^N \lambda_j(1-z_j))$, whereas the remaining $L_i^{(behind)}$ customers have not been served, and hence, are "effectively replaced" by a population whose PGF equals z_i.

Approximations: The results presented in Theorems 4 and 5 suggest the following simple approximations for the waiting-time distributions for stable systems: For $\rho < 1$, $i = 1, \ldots, N$,

$$\Pr\{W_i < x\} \approx \Pr\{\tilde{W}_i < x(1-\rho)\}, \tag{97}$$

and similarly for the moments: for $\rho < 1$, $i = 1, \ldots, N$, $k = 1, 2 \ldots$,

$$E[W_i^k] \approx \frac{E[\tilde{W}_i^k]}{(1-\rho)^k}, \tag{98}$$

where closed-form expressions $E[\tilde{W}_i^{(k)}]$ can be directly obtained from Theorems 4 and 5 by k-fold differentiation. Extensive validation of these appoximations fall beyond the scope of this paper. We refer to [36, 40, 42] for extensive discussions about the accuracy of these approximations for the special case of exhaustive and gated service.

4 Topics for Further Research

The results presented in this paper provide a significant step towards the development of a unified theory of polling in heavy traffic. Nonetheless, the results raise a number of challenging open questions for further research. First, in this paper it is assumed that the second moments of the service-time distributions are finite, forced by the second-moment assumption on the offspring function, needed for the validity of Theorem 1. An interesting area for further research is to obtain heavy-traffic results for heavy-tailed service-time distributions with infinite variance. In this context, interesting results have been obtained by Boxma et al. [5], who study the tail behavior of the waiting times in polling systems with so-called regularly varying service times and switch-over times, and by Boxma and Cohen [7], who derive the heavy-traffic limiting distribution for the waiting times in the single-server queue with heavy-tailed service-time distributions. Second, in order to use the theory of MTBPs the arrival processes must be Poisson (or batched Poisson). Interestingly, in special cases similar heavy-traffic results have been obtained under the weaker assumption of independent renewal processes, where also the gamma-distribution appears to play a key role (see for example [12, 44]). Note, however, that the proofs of these results for $N > 2$ are based on partial conjectures. Moreover, for several polling models it was found that the heavy-traffic limits of a Poisson-type model and its renewal counterpart only differ by a simple scaling constant (see for example [43, 44] for non-cyclic

asymptotic behavior of polling models in the general setting of the present paper, with renewal arrivals. Finally, a related area of of research is the analysis of the waiting times in polling systems with multiple (say $m > 1$) servers. Multiple-server polling models are notoriously hard, and do not leave any hope for an exact analysis. Interestingly, based on numerical experimentation it was obseved in [24, 9, 41] that if the servers follows the same route they tend to cluster together, particularly when the system is heavily loaded. These results suggest that in the limiting case all servers tend to effectively work as a single server that works m times as fast. This, in turn, suggests that we may use our heavy-traffic results for single-server polling models to develop simple approximations for the delay figures at each of the queues. Preliminary experimentation with simulations show promising results, opening up an interesting area for further research.

Acknowledgment: The author wishes to thank Ton Dieker for his useful suggestions.

References

[1] Altman, E., Khamisy, A. and Yechiali, U. (1992). On elevator polling with globally gated regime. *Queueing Systems* **11**, 85-90.

[2] Athreya, K.B. and Ney, P.E. (1972). *Branching Processes* (Springer, Berlin).

[3] Blanc, J.P.C. (1992). An algorithmic solution of polling systems with limited service disciplines. *IEEE Trans. Commun.* **40**, 1152-1155.

[4] Blanc, J.P.C. (1992). Performance evaluation of polling systems by means of the power-series algorithm. *Ann. Oper. Res.* **35**, 155-186.

[5] Boxma, O.J., Deng, Q., and Resing, J.A.C. (2000). Polling systems with regularly varying service and/or switchover times. *Adv. Perf. Anal.* **3**, 71-107.

[6] Boxma, O.J., Levy, H. and Yechiali, U. (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Ann. Oper. Res.* **35**, 187-208.

[7] Boxma, O.J. and Cohen, J.W. (1999). Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions. *Queueing Systems* **33**, 177-204.

[8] Borst, S.C. and Boxma, O.J. (1997). Polling systems with and without swicth-over times. *Oper. Res.* **45**, 536-543.

[9] Borst, S.C. and Van der Mei, R.D. (1999). Waiting-time approximations for multiple-server polling systems. *Perf. Eval.* **31**, 163-182.

[10] Choudhury, G. and Whitt, W. (1996). Computing transient and steady state distributions in polling models by numerical transform inversion. *Perf. Eval.* **25**, 267-292.

[11] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1995). Polling systems with zero switch-over times: a heavy-traffic principle. *Ann. Appl. Prob.* **5**, 681-719.

[12] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1998). Polling systems in heavy-traffic: a Bessel process limit. *Math. Oper. Res.* **23**, 257-304.

[13] Fricker, C. and Jaïbi, M.R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* **15**, 211-238.

[14] Foster, J.H. (1969). *Branching Processes Involving Immigration.* Ph.D. Thesis, University of Wisconsin. (A hard copy is available from the author upon request.)

[15] Joffe, A. and Spitzer, F. (1967). On multitype branching processes with $\rho \leq 1$. *Math. Anal. Appl.* **19**, 409-430.

[16] Konheim, A.G., Levy, H. and Srinivasan, M.M. (1994). Descendant set: an efficient approach for the analysis of polling systems. *IEEE Trans. Commun.* **42**, 1245-1253.

[17] Kroese, D.P. (1997). Heavy traffic analysis for continuous polling models. *J. Appl. Prob.* **34**, 720-732.

[18] Kudoh, S., Takagi, H. and Hashida, O. (2000). Second moments of the waiting time in symmetric polling systems. *J. Oper. Res. Soc. of Japan* **43**, 306-316.

Distributed and parallel Systems, eds. T. Hasegawa, H. Takagi and Y. Takahashi (North-Holland, Amsterdam), 127-139.

[20] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750–1760.

[21] Levy, H. and Sidi, M. (1991). Polling systems with simultaneous arrivals. *IEEE Trans. Commun.* **39**, 823-827.

[22] Markowitz, D. and Wein, L.M. (2001). Heavy traffic analysis of dynamic cyclic policies: a unified treatment of the single machine scheduling problem. *Oper. Res.* **49**, 246-270.

[23] Markowitz, D., Reiman, M.I. and Wein, L.M. (2000). The stochastic economic lot scheduling problem: heavy traffic analysis of dynamic cyclic policies. *Oper. Res.* **48**, 136-154.

[24] Morris, R.J.T. and Wang, Y.T. (1984). Some results for multi-queue systems with multiple cyclic servers. In: *Performance of Computer Communication Systems*, eds. W. Bux and H. Rudin (North-Holland, Amsterdam), 245-258.

[25] Park, C.G., Han, D.H., Kim, B. and Jun, H.-S. (2005). Queueing analysis of symmetric polling algorithm for DBA scheme in an EPON. In: *Proc. Korea-Netherlands joint conference on Queueing Theory and its Applications to Telecommunication Systems*, ed. B.D. Choi (Seoul, June 22-25), 147-154.

[26] Quine, M.P. (1972). The multitype Galton-Watson process with ρ near 1. *Adv. Appl. Prob.* **4**, 429-452.

[27] Reiman, M.I. and Wein, L.M. (1998). Dynamic scheduling of a two-class queue with setups. *Oper. Res.* **46**, 532-547.

[28] Reiman, M.I., Rubio, R. and Wein, L.M. (1999). Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transp. Sc.* **33**, 361-380.

[29] Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* **13**, 409-426.

[30] Shimogawa, S. and Takahashi, Y. (1992). A note on the conservation law for a multi-queue with local priority. *Queueing Systems* **11**, 145-151.

[31] Sidi, M. and Levy, H. (1990). Customer routing in polling systems. In: *Proc. Performance '90*, eds. P.J.B. King, I. Mitrani and R.B. Pooley (North-Holland, Amsterdam), 319-331.

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

[33] Takagi, H. (1990). Queueing analysis of polling models: an update. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 267-318.

[34] Takagi, H. (1997). Queueing analysis of polling models: progress in 1990-1994. In: *Frontiers in Queueing: Models and Applications in Science and Technology*, ed. J.H. Dshalalow (CRC Press, Boca Raton, FL), 119-146.

[35] Takagi, H. (1991). Application of polling models to computer networks. *Comp. Netw. ISDN Syst.* **22**, 193-211.

[36] Van der Mei, R.D. (1999). Distribution of the delay in polling systems in heavy traffic. *Perf. Eval.* **31**, 163-182.

[37] Van der Mei, R.D. (1999). Polling systems in heavy traffic: higher moments of the delay. *Queueing Systems* **31**, 265-294.

[38] Van der Mei, R.D. (2000). Polling systems with switch-over times under heavy load: moments of the delay. *Queueing Systems* **36**, 381-404.

[39] Van der Mei, R.D. (2002). Waiting-time distributions in polling systems with simultaneous batch arrivals. *Ann. Oper. Res.* **113**, 157-173.

[40] Van der Mei, R.D. and Levy, H. (1998). Expected delay analysis in polling systems in heavy traffic. *J. Appl. Prob.* **30**, 586-602.

power-series algorithm. *Stoch. Mod.* **13**, 339-363.

[42] Van der Mei, R.D. and Levy, H. (1997). Polling systems in heavy traffic: exhaustiveness of service policies. *Queueing Systems* **27**, 227-250.

[43] Olsen, T.L. and Van der Mei, R.D. (2003). Periodic polling systems in heavy-traffic: distribution of the delay. *J. Appl. Prob.* **40**, 305-326.

[44] Olsen, T.L. and Van der Mei, R.D. (2005). Periodic polling systems in heavy-traffic: renewal arrivals. *Oper. Res. Lett.* **33**, 17-25.

[45] Vatutin, V.A. and Dyakonova, E.E. (2002). Multitype branching processes ans some queueing systems. *J. of Math. Sciences* **111**, 3901-3909.

[46] Vishnevskii, V.M. and Semenova, O.V. (2006). Mathematical methods to study the polling systems. *Automation and Remote Control* **67**, 173-220.

Appendix A: Proof of Lemma 3

Part 1 was shown in [29]. Part 2 follows from the fact that all entries of $\mathbf{M}$ are continuous functions of ρ, which implies that continuity of $\xi = \xi(\rho)$ with respect to ρ, which implies the continuity of $\xi(\rho)$ with respect to ρ (see for example [2]). The fact that $\xi(1) = 1$ follows directly from the fact that $\hat{\mathbf{M}}\underline{b} = \underline{b}$, which is an immediate consequence of the fact that the GG-model described in Section 3.1 is work conserving. Finally, to prove Part 4 we adopt the concept and notation of the Descendant Set Approach (DSA) from [16]. The DSA focuses on an arbitrary polling instant of the server at Q_1, called the reference point, and focuses on X_1, the number of type-1 customers in the system at that moment. Denoting by $A_{i,c}$ the contribution to X_1 of a type-i customer that was present in the system at a polling instant c cycles before the reference point, the mean values $\alpha_{i,c} := E[A_{i,c}]$ can be obtained via the following recursive relations (cf. [16] for details): For $i = 1, \ldots, N$,

$$\alpha_{i,-1} := I_{\{i=1\}}, \tag{99}$$

and for $c = 0, 1, \ldots$,

$$\alpha_{i,c} = b_i^{(1)} \sum_{j=1}^{N} \lambda_j \alpha_{j,c-1}. \tag{100}$$

Then if we define, for $\rho < 1$,

$$\Delta := \sum_{i=1}^{N} \lambda_i \sum_{c=0}^{\infty} \alpha_{i,c}, \tag{101}$$

then substitution of (99) and (100) immediately leads to the observation that, for $\rho < 1$,

$$\Delta = \rho\,(\Delta + \lambda_1) = \frac{\lambda_1 \rho}{1-\rho}. \tag{102}$$

Alternatively, based on known properties for the maximum eigenvalue we can decompose $\alpha_{i,c}$ into a dominant and a recessive part as follows (see for example [2, 40]): for $\rho < 1$,

$$\alpha_{i,c} = \xi^{c+1} w_i v_1 + s_{i,c} \tag{103}$$

where $s_{i,c}$ is a lower-order term in the sense that there exists K $(0 < K < \infty)$ and ξ_* $(0 < \xi_* < \xi)$ such that $|s_{i,c}| < K\xi_*^c$ for all $c = 0, 1, \ldots$, which is readily seen to imply that, for $i = 1, \ldots, N$,

$$\sum_{c=0}^{\infty} s_{i,c} < \infty. \tag{104}$$

From (103) we have, for $\rho < 1$ (and hence $\xi < 1$, see part 1 of Lemma 3),

$$\Delta = \frac{\xi}{1-\xi} \sum_{i=1}^{N} \lambda_i w_i v_1 + \sum_{i=1}^{N} \lambda_i \sum_{c=0}^{\infty} s_{i,c}. \tag{105}$$

(104) and parts 1, 2 and 3 of Lemma 3 we obtain

$$\hat{\lambda}_1\hat{\rho} = \hat{\xi}\hat{\rho}\hat{\lambda}_1 \lim_{\rho\uparrow 1}\frac{1-\rho}{1-\xi} + 0 = \hat{\lambda}_1 \lim_{\rho\uparrow 1}\frac{1-\rho}{1-\xi}, \tag{106}$$

which immediately implies

$$\lim_{\rho\uparrow 1}\ \frac{1-\xi}{1-\rho} = 1. \tag{107}$$

This completes the proof of Lemma 3. □

Appendix B: Proof of Lemma 4

Parts 1, 2 and 3 follow from similar arguments as those of Lemma 3. To prove Part 4, for the cyclic branching model the Desendant Set variables $\alpha_{i,c}$ (defined above) satisfy the following recursive equations (cf. also [40]): For $i = 1, \ldots, N$, $\alpha_{i,-1} := I_{\{i=1\}}$, and for $c = 0, 1, \ldots$,

$$\alpha_{i,c} = E[T_i]\left(\sum_{j=i+1}^{N}\lambda_j\alpha_{i,c} + \sum_{j=1}^{i-1}\lambda_j\alpha_{j,c-1}\right) + E[L_i]\alpha_{j,c-1} \tag{108}$$

$$= f_i\frac{b_i^{(1)}}{1-\rho_i}\left(\sum_{j=i+1}^{N}\lambda_j\alpha_{j,c} + \sum_{j=1}^{i-1}\lambda_j\alpha_{j,c-1}\right) + (1-f_i)\alpha_{i,c-1}. \tag{109}$$

Then if Δ is defined as in (101) it is readily verified that, for $\rho < 1$ (and hence also $\xi < 1$, see part 1 of Lemma 4),

$$\Delta = \frac{\lambda_1(1-\rho_1 - f_1(1-\rho))}{f_1(1-\rho)}. \tag{110}$$

Then similar to the proof of Lemma 3 above we can write, for $\rho < 1$,

$$\alpha_{i,c} = \xi^{c+1}w_iv_1 + s_{i,c}, \tag{111}$$

where $s_{i,c}$ satisfies (104). This implies that, for $\rho < 1$,

$$\Delta = \frac{\xi}{1-\xi}v_1\sum_{i=1}^{N}\lambda_iw_i + \sum_{i=1}^{N}\lambda_i\sum_{c=0}^{\infty}s_{i,c}. \tag{112}$$

Then following similar arguments as in the proof of Lemma 3 in Appendix A, combining (64), (94), (110), (112) and parts 1, 2 and 3 of Lemma 4 we obtain

$$\lim_{\rho\uparrow 1}\ \frac{1-\xi}{1-\rho} = \frac{1}{\delta}. \tag{113}$$

This completes the proof of Lemma 4. □