On a Unifying Theory on Polling Models in Heavy Traffic

R.D. van der Mei

CWI, Probability and Stochastic Networks, Amsterdam, The Netherlands
Vrije Universiteit, Faculty of Sciences, Amsterdam, The Netherlands
mei@few.vu.nl

Abstract. 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 $\underline{X}$ that describes the state of the system at these polling instants satisfies the following heavy-traffic behavior, under mild assumptions:

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

where $\underline{\gamma}$ is a known 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. As an illustration of the usefulness of the results, we derive new closed-form expressions for the LST of the waiting-time distributions for models with a cyclic globally-gated polling regime. 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 models, queueing theory, heavy traffic, unfication.

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;

L. Mason, T. Drwiega, and J. Yan (Eds.): ITC 2007, LNCS 4516, pp. 556–567, 2007.

the reader is referred to [24,13] 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 [22,23], and to a recent survey paper by Vishnevskii and Semenova [34] 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 [18] 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. [6,7], 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 [25], periodic server routing [31,32], simultaneous batch arrivals [28], continuous polling [11], amongst others. In this context, a remarkable observation is that in the heavy-traffic behavior of polling models a central role is played by the gamma-distribution, which occurs in the analysis of these different model variants as the 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 (e.g., [12]). 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. 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 [5] and non-branching-type polling models [3], which raises the need for simple and fast approximations. To this end, heavy-traffic asymptotics form an excellent basis for developing such approximations,

and in fact, have been found to be remarkably accurate in several cases, even for moderate load [25,27,32].

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 largely developed in the early 1970s, is well-matured and powerful [17,9]. 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 Quine [17]:

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

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 heavy-traffic asymptotics for polling model that have MTBP-structure. To demonstrate the usefulness of the 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 [17] (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 MTBP-type polling models. To illustrate the usefulness of the approach, we derive closed-form expressions for the LST of the scaled asymptotic waiting-time distributions for cyclic models with GG service. The implications of these results are discussed extensively.

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 $|z_k| \le 1$ $(k = 1,\ldots,M), i = 1,\ldots,M$,

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

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 $|z_k| \le 1$ $(k = 1,\ldots,M)$,

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

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 [2], normalized such that

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

The following conditions are necessary and sufficient conditions for the ergodicity of the process $\mathbf{Z}$ (cf. [18]): $\xi < 1$ 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 was shown in [17]:

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

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

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. As an illustration, we consider a cyclic polling model with globally-gated (GG) service, in a general parameter setting. For this model, we show how Theorem 1 can be used to obtain derive the LST of the asymptotic waiting-time distribution at each of the queues. In Section 3.1 we discribe the GG-model, in Section 3.2 we derive HT-asymptotics for the waiting-time distributions at each of the queues and in Section 3.3 we extensively discuss the implications of the results.

3.1 Model

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 is $\rho := \sum_{i=1}^{N} \rho_i$. We define a polling instant at Q_i to be the moment at which the server arrives at Q_i, and a departure epoch at Q_i a moment at which the server depart from Q_i. The 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. The GG service discipline works as follows (cf. [4]). 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. [8,4]). Throughout this paper, this model will be referred to as the GG-model. 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 GG-model model the 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.

3.2 Analysis

To analyze the HT-behavior of the GG-model, we proceed along a number of steps. First, we establish the relation between the GG-model and the general MTBP-model described in Section 2. Then, we use Theorem 1 to obtain HT-limits for the joint queue-length vector at polling instants at a fixed queue (Theorem 2). Finally, this result is transformed into an expression for the asymptotic scaled waiting-time distribution at an arbitrary queue (Theorem 3). For compactness of the presentation, the proofs of the various results are omitted.

To start, we consider the MTBP $\mathbf{X}^{(1)} := \{\underline{X}_n^{(1)}, n = 0, 1, \ldots\}$ describing the evolution of the joint queue-length 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

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

Note that it follows directly from (16) that, for $j = 1, \ldots, N$,

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

To derive the limiting distribution of the joint queue-length vector at polling instants at Q_1, we need to specify the following parameters: (a) 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 (b) 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 $\hat{\mathbf{M}}$ *is given by the following expression:*

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

Moreover, the right and left eigenvectors of $\hat{\mathbf{M}}$ *are*

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

with

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

Lemma 2
For the GG-model, we have

$$\underline{\hat{g}}^\top \underline{\hat{w}} = |\underline{b}|^{-1} r, \quad \textit{and} \quad A = |\underline{b}|^{-1} \frac{b^{(2)}}{b^{(1)}}. \tag{21}$$

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

$$\xi = \xi(\rho), \tag{22}$$

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*

$$\xi'(1) := \lim_{\rho \uparrow 1} \frac{1 - \xi(\rho)}{1 - \rho} = 1. \tag{23}$$

The following result transform Theorem 1 into an HT-result for the GG-model under consideration.

Theorem 2
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:

$$(1-\rho)\begin{pmatrix} X_1^{(k)} \\ \vdots \\ X_N^{(k)} \end{pmatrix} \to_d \frac{b^{(2)}}{b^{(1)}} \left[(\hat{\rho}_1 + \cdots + \hat{\rho}_{k-1}) \begin{pmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_{k-1} \\ \hat{\lambda}_k \\ \vdots \\ \hat{\lambda}_N \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \hat{\lambda}_k \\ \vdots \\ \hat{\lambda}_N \end{pmatrix} \right] \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{24}$$

where

$$\alpha = r\frac{b^{(1)}}{b^{(2)}}. \tag{25}$$

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

Theorem 3
For the GG-model, the waiting-time distribution satisfies the following limiting behavior: For $i = 1, \ldots, N$,

$$(1-\rho)W_i \to_d \tilde{W}_i \quad (\rho \uparrow 1) \tag{26}$$

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

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

where

$$\alpha = r\frac{b^{(1)}}{b^{(2)}}, \text{ and } \mu = \frac{b^{(1)}}{b^{(2)}}. \tag{27}$$

3.3 Discussion

Theorem 3 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 $b^{(2)}$,*
(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.

We end this session with a number of remarks.

Remark 1 (Model extensions): The results presented for the GG-model described in Section 3.1 mainly serve as an illustration, and can be readily extended to a broader set of models. The requirements for the derivation of heavy-traffic limits similar to Theorems 2 and 3 are that (a) the evolution of the system at specific moments can be described as a multi-dimensional branching process with immigration, and (b) that the system 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 [31], models with (simultaneous) batch arrivals [28,14], continuous polling models [11], models with customer routing [20], globally-gated models with elevator-type routing [1], models with local priorities [19], 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.

Remark 2 (Assumptions on the finiteness of moments): Theorems 2 and 3 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 [25] 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 [25] 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 [27,26]. 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 Theorem 3.

Remark 3 (Approximations): The results presented in Theorem 3 suggest the following simple approximations for the waiting-time distributions for stable systems: For $\rho < 1$, $i = 1, \ldots, N$,

$$\Pr\{W_i < x\} \approx \Pr\{\tilde{W}_i < x(1-\rho)\}, \tag{28}$$

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

$$E[W_i^k] \approx \frac{E[\tilde{W}_i^k]}{(1-\rho)^k}, \tag{29}$$

where closed-form expressions for $E[\tilde{W}_i^k]$ can be directly obtained from Theorem 3 by k-fold differentiation. Extensive validation of these appoximations fall beyond the scope of this paper. We refer to [25,29,30] for extensive discussions about the accuracy of these approximations for the special case of exhaustive and gated service.

References

1. Altman, E., Khamisy, A., Yechiali, U.: On elevator polling with globally gated regime. Queueing Systems 11, 85–90 (1992)
2. Athreya, K.B., Ney, P.E.: Branching Processes. Springer, Berlin (1972)
3. Blanc, J.P.C.: Performance evaluation of polling systems by means of the power-series algorithm. Ann. Oper. Res. 35, 155–186 (1992)
4. Boxma, O.J., Levy, H., Yechiali, U.: Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. Ann. Oper. Res. 35, 187–208 (1992)
5. Choudhury, G., Whitt, W.: Computing transient and steady state distributions in polling models by numerical transform inversion. Perf. Eval. 25, 267–292 (1996)
6. Coffman, E.G., Puhalskii, A.A., Reiman, M.I.: Polling systems with zero switch-over times: a heavy-traffic principle. Ann. Appl. Prob. 5, 681–719 (1995)
7. Coffman, E.G., Puhalskii, A.A., Reiman, M.I.: Polling systems in heavy-traffic: a Bessel process limit. Math. Oper. Res. 23, 257–304 (1998)
8. Fricker, C., Jaïbi, M.R.: Monotonicity and stability of periodic polling models. Queueing Systems 15, 211–238 (1994)
9. Joffe, A., Spitzer, F.: On multitype branching processes with $\rho \leq 1$. Math. Anal. Appl. 19, 409–430 (1967)
10. Konheim, A.G., Levy, H., Srinivasan, M.M.: Descendant set: an efficient approach for the analysis of polling systems. IEEE Trans. Commun. 42, 1245–1253 (1994)
11. Kroese, D.P.: Heavy traffic analysis for continuous polling models. J. Appl. Prob. 34, 720–732 (1997)
12. Kudoh, S., Takagi, H., Hashida, O.: Second moments of the waiting time in symmetric polling systems. J. Oper. Res. Soc. of Japan 43, 306–316 (2000)
13. Levy, H., Sidi, M.: Polling models: applications, modeling and optimization. IEEE Trans. Commun. 38, 1750–1760 (1991)
14. Levy, H., Sidi, M.: Polling systems with simultaneous arrivals. IEEE Trans. Commun. 39, 823–827 (1991)
15. Morris, R.J.T., Wang, Y.T.: Some results for multi-queue systems with multiple cyclic servers. In: Bux, W., Rudin, H. (eds.) Performance of Computer Communication Systems, North-Holland, Amsterdam, pp. 245–258 (1984)
16. Park, C.G., Han, D.H., Kim, B., Jun, H.-S.: Queueing analysis of symmetric polling algorithm for DBA scheme in an EPON. In: Choi, B.D. (ed.) Proc. Korea-Netherlands joint conference on Queueing Theory and its Applications to Telecommunication Systems, Seoul, pp. 147–154 (June 22-25, 2005)
17. Quine, M.P.: The multitype Galton-Watson process with ρ near 1. Adv. Appl. Prob. 4, 429–452 (1972)

18. Resing, J.A.C.: Polling systems and multitype branching processes. Queueing Systems 13, 409–426 (1993)
19. Shimogawa, S., Takahashi, Y.: A note on the conservation law for a multi-queue with local priority. Queueing Systems 11, 145–151 (1992)
20. Sidi, M., Levy, H.: Customer routing in polling systems. In: King, P.J.B., Mitrani, I., Pooley, R.B., (eds.) Proc. Performance '90, North-Holland, Amsterdam, pp. 319–331(1990)
21. Takagi, H.: Analysis of Polling Systems. MIT Press, Cambridge, MA (1986)
22. Takagi, H.: Queueing analysis of polling models: an update. In: Takagi, H., (ed.) Stochastic Analysis of Computer and Communication Systems, North-Holland, Amsterdam, pp. 267–318 (1990)
23. Takagi, H.: Queueing analysis of polling models: progress in 1990-1994. In: Dshalalow, J.H. (ed.) Frontiers in Queueing: Models and Applications in Science and Technology, pp. 119–146. CRC Press, Boca Raton, FL (1997)
24. Takagi, H.: Application of polling models to computer networks. Comp. Netw. ISDN Syst. 22, 193–211 (1991)
25. Van der Mei, R.D.: Distribution of the delay in polling systems in heavy traffic. Perf. Eval. 31, 163–182 (1999)
26. Van der Mei, R.D.: Polling systems in heavy traffic: higher moments of the delay. Queueing Systems 31, 265–294 (1999)
27. Van der Mei, R.D.: Polling systems with switch-over times under heavy load: moments of the delay. Queueing Systems 36, 381–404 (2000)
28. Van der Mei, R.D.: Waiting-time distributions in polling systems with simultaneous batch arrivals. Ann. Oper. Res. 113, 157–173 (2002)
29. Van der Mei, R.D., Levy, H.: Expected delay analysis in polling systems in heavy traffic. J. Appl. Prob. 30, 586–602 (1998)
30. Van der Mei, R.D., Levy, H.: Polling systems in heavy traffic: exhaustiveness of service policies. Queueing Systems 27, 227–250 (1997)
31. Olsen, T.L., Van der Mei, R.D.: Periodic polling systems in heavy-traffic: distribution of the delay. J. Appl. Prob. 40, 305–326 (2003)
32. Olsen, T.L., Van der Mei, R.D.: Periodic polling systems in heavy-traffic: renewal arrivals. Oper. Res. Lett. 33, 17–25 (2005)
33. Vatutin, V.A., Dyakonova, E.E.: Multitype branching processes ans some queueing systems. J. of Math. Sciences 111, 3901–3909 (2002)
34. Vishnevskii, V.M., Semenova, O.V.: Mathematical methods to study the polling systems. Automation and Remote Control 67, 173–220 (2006)