Polling Systems with Two-Phase Gated Service: Heavy Traffic Results for the Waiting Time Distribution

R.D. van der Mei[a,b] and J.A.C. Resing[c]

[a]Centre for Mathematics and Computer Science
Department of Probability and Stochastic Networks
Amsterdam, Netherlands

[b]Vrije Universiteit
Faculty of Sciences, Department of Mathematics
Amsterdam, Netherlands

[c]Eindhoven University of Technology
Department of Mathematics and Computer Science
Eindhoven, The Netherlands

E-mail: mei@cwi.nl, resing@win.tue.nl

Abstract

We study an asymmetric cyclic polling system with Poisson arrivals, general service-time and switch-over time distributions, and with so-called two-phase gated service at each queue, an interleaving scheme that aims to enforce some level of "fairness" among the different customer classes. For this model, we use the classical theory of multi-type branching processes (MTBPs) to derive closed-form expressions for the Laplace-Stieltjes Transform (LST) of the waiting-time distributions when the load tends to 1, in a general parameter setting and under proper heavy-traffic (HT) scalings. This result is strikingly simple and provides new insights in the behavior of two-phase polling systems. In particular, the result provides insight in the waiting-time performance, and the tradeoff between efficiency and fairness of two-phase gated polling compared to the classical one-phase gated service policy.

1 Introduction

A polling system is a multi-queue single-server system in which the server visits the queues in cyclic order to process requests pending at the queues. Polling models occur naturally in the modeling of systems in which service capacity (e.g., CPU, bandwidth) is shared by different types of users, each type having specific traffic characteristics and performance requirements. Polling models find a variety of applications in the areas of computer-communication networks, production, manufacturing and maintenance [17, 31]. In many applications of polling models a key issue is how to realize some level of fairness among different customer classes. Motivated by this, many service disciplines have been proposed containing some kind of interleaving scheme to enforce fairness among different customer classes by somehow limiting the number of customers served during a single visit of the server to a queue. This has led to the definition of a variety of service policies, including for example the classical K-limited, time-limited or Bernoulli-type service policies which

put (either fixed or stochastic) upper bounds to the duration of a visit of the server to a queue. In addition, a number of variants of exhaustive and gated service policies have been proposed to avoid monopolization by a single queue, including for example the binomial-gated, fractional-exhaustive or Bernoulli-type service policies, amongst others. The two-phase gated policy analyzed in this paper may be viewed as an interesting alternative to the classical gated service policy.

The motivation for studying polling models with two-phase gated service is twofold. First, the ultimate goal of studying the performance of polling models is to understand how to efficiently operate the system, e.g. in terms of 'How many customers should be served during a visit of the server to a queue?' and 'In what order should the queues be visited by the server?' In this context, two important issues are often considered: *efficiency* and *fairness*. The two-phase gated service policy provides a promising means to realize fairness by enforcing some level of interleaving between differentent customer classes. As such two-phase gated service may be seen as an interesting alternative to the classical limited-type (e.g., Bernoulli, K-limited, time-limited) or fractional-type (e.g., binomial-gated, fractional-exhaustive) service policies. In this context, the aim of his paper is to quantify the trade-off between efficiency and fairness by comparing the two-phase gated model with the classical one-phase gated model. Second, two-phase service policies find specific applications in the area of Ethernet Passive Optical Networks (EPONs). In fact, Kramer et al. [14] propose two-phase scheduling policies to implement a dynamic bandwidth allocation scheme in an Ethernet Passive Optical network (EPON), where packets from different Optical Network Units (ONUs) share channel capacity in the upstream direction. An EPON is a point-to-multipoint network in the downstream direction and a multi-point to point network in the upstream direction. The Optical Line Terminal (OLT) resides in the local office, connecting the access network to the Internet. The OLT allocates the bandwidth to the Optical Network Units (ONUs) located at the customer premises, providing interfaces between the OLT and end-user network to send vocie, video and data traffic. In an EPON the process of transmitting data downstream from the OLT to the ONUs is broadcast in variable-length packet according to the 802.3 protocol [12]. However, in the upstream direction the ONUs share capacity, and various polling-based bandwidth allocation schemes can be implemented. Simple time-division multiplexing access (TDMA) schemes based on fixed time-slot assignment suffer from the lack of statistical multiplexing, making inefficient use of the available bandwidth, which raises the need for dynamic bandwidth allocation (DBA) schemes. A dynamic scheme that reduces the time-slot size when there are no data to transmit would allow excess bandwidth to be used by other ONUs. However, the main obstacle of implementing such a scheme is the fact the OLT does not know in advance how much data each ONU has to transmit. To overcome this problem, Kramer et al. [12, 13, 14] propose an OLT-based interleaved polling scheme similar to hub-polling to support dynamic bandwidth allocation. To avoid monopolization of bandwidth usage of ONUs with high data volumes they propose an interleaved DBA scheme with a maximum transmission window size limit. Motivated by this, Park et al. [20] proposed the two-phase gated service as a means to enforce interleaving between different traffic streams, aiming to realize some degree of "fairness" amongst different customer classes. To this end, they derive a pseudo-conservation law for the two-phase gated system and use the classical buffer-occupancy method to express the expected delay as the solution of a set of linear equations. We believe that apart from its application in the specific context of EPONs, the two-phase service policy may also be an interesting "fair" alternative to the classical gated service policy in other application areas, such as production, manufacturing and maintenance.

Despite the fact that fairness is an important aspect in queueing models, there is no commonly accepted notion of fairness in queueing systems and how to quantify it. Wierman and Harchol-Balter [44] propose to use as a fairness criterion the $E[T(x)]/x$, where $T(x)$ stands for the conditional sojourn time of a job of size x, to evaluate whether a system is fair or unfair. Raz et al. [25] go a step further by proposing the so-called Resource Allocation Queueing Fairness Measure (RAQFM) as an unfairness *measure* that takes into account both *seniority* and *service-time differences*; they also show that queue fairness is sensitive to service-time variability and that the fairness ranking of commonly used scheduling policies (such as FCFS, LCFS, ROS) depends on this parameter.

Recently, Brosh et al. [6] have proposed the so-called Slowdown Queueing Fairness (SQF) measure, based on a proportionality principle as the underlying belief. SQF can be viewed as bridging the gap between the slowdown expected criterion [44] (which focuses on service requirements only, not on seniority) and the "natural" waiting-time variance (which focuses on job seniority, not on service requirements). We refer to [6] for a recent overview of the available literature on fairness in queueing systems. To the best of the authors' knowledge fairness has not been addressed in the context of polling models before.

The analysis of polling models has received much attention over the past couple of decades. One of the most remarkable results is that there appears to be a striking difference in complexity between polling models. Resing [28] observed that for a large class of polling models, including for example cyclic or periodic polling models with Poisson arrivals, exhaustive or gated service at all queues, and switch-over times that are independent of the state of the system, 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 and require heavy-weight numerical techniques to obtain the queue-length and waiting-time distributions [4, 5]. For MTBP-type polling models a number of solution techniques have been proposed, including for example the classical buffer-occupancy and station-time techniques [29], the Descendant Set Approach [11] and the recently proposed Mean Value Analysis [45]. As an interesting extension of MTBP-type models, Groenevelt and Altman [10] and Altman and Fiems [1] consider polling models in which the switch-over times are correlated, by using stochastic recursive equations. Their results show that the correlations between the switch-over times may have a significant impact on the waiting-time performance of the system. We refer to [29, 30, 32, 43] for overviews of the available results on polling models.

There are several strong reasons for considering heavy traffic (HT) asymptotics. Exact analysis of the delay in polling models is only possible in some cases, and even in those cases numerical techniques are usually required to obtain the expected delay at each of the queues. However, the use of numerical techniques for the analysis of polling models has several drawbacks. First, numerical techniques do not reveal explicitly how the system performance depends on the system parameters and can therefore contribute to the understanding of the system behavior only to a limited extent. Exact closed-form expressions provide much more insight into the dependence of the performance measures on the system parameters. Second, the efficiency of each of the numerical algorithms degrades significantly for heavily loaded, highly asymmetric systems with a large number of queues, while the proper operation of the system is particularly critical when the system is heavily loaded. These observations raise the importance of an exact asymptotic analysis of the delay in polling models in HT.

Over the past decade, polling models in HT have received significant attention. For a two-queue model with exhaustive service at both queues, Coffman et al. [7, 8] use an averaging principle to derive expressions for the workload and waiting-time distributions under HT assumptions. For models with independent Poisson arrivals, Kudoh et al. [16] 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 [30]. They also give conjectures for the HT 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) [11] to derive exact expressions for the waiting-time distributions in models with mixtures of exhaustive and gated service and cyclic [33] or periodic [21] server routing, and with simultaneous batch arrivals [36]. Van der Mei [37] considers the general class of polling models that can be described by MTBPs [28] and uses the theory of critical MTBP [24] to obtain a framework for deriving HT-limits for the waiting-time and queue-length disributions. Van der Mei and Winands [41] use the Mean Value Analysis (MVA) [45] to derive HT limits for the expected delay for cyclic Poisson-driven

polling models with exhaustive and gated service at all queues. In [40] they derive expressions for the expected delay in cyclic polling models with gated and exhaustive service, providing rigorous proofs for the results conjectured earlier in [7, 22]. Kroese [15] studies continuous polling systems in HT with unit renewal arrivals on a circle and shows that the steady-state number of customers has approximately a gamma-distribution. Vatutin and Dyakonova [42] use the theory of MTBPs to obtain the limiting distributions for several two-queue polling models with zero switch-over times. Altman and Kushner [1] study the HT-behavior of polling models in which the queue may be temporarily unavailable. For this model, they show that the suitably scaled total workloads converge to a controlled limit diffusion process with jumps. They also show that the individual queued workloads and job numbers can be recovered (asymptotically) from the limiting scaled workload. Another interesting limiting regime in which the queue lengths grow to infinity is when the switch-over times are large. In this case, strikingly simple results about the distributions of the delay can be obtained [23, 46, 34]. 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 [18, 19, 26, 27] and references therein.

We consider an asymmetric cyclic polling model with N queues and with generally distributed service times and switch-over times. Each queue receives so-called two-phase gated service, which works as follows: Newly incoming customers are first queued at the phase-1 buffer. When the server arrives at a queue, it closes the gate behind the customers residing in the phase-1 buffer, then serves all customers waiting in the phase-2 buffer on a FCFS basis, and moves all customers before the gate at the phase-1 buffer to the phase-2 buffer before moving to the next queue. In a recent paper [39] we studied the *mean* of the delay W_i incurred at queue i, when the load ρ tends to unity, under proper HT scalings. Amongst others, the results in [39] explicitly quantify the trade-off between the decrease of efficiency and the increase in the so-called *queue fairness* introduced by implementing the two-phase gated service policy (in comparison with the classical one-phase gated service policy); queue fairness is a simple fairness measure that only depends on the expected delay at each of the queues. An interesting alternative notion of fairness is *customer fairness*, which considers fairness between individual customers (e.g., depending on their seniority and service-time requirements). Recently, Brosh et al. [6] proposed the so-called Slowdown Queueing Fairness (SQF) as a customer fairness measure. However, to quantify the SQF of a polling model, we need to quantify the *second moments* of the waiting times at the queues. Motivated by this, in this paper we focus on the complete *distribution* of W_i when ρ tends to unity, under proper HT scalings. Following the general lines discussed in [37] for the derivation of HT-asymptotics for branching-type polling models, we obtain a closed-form expression for the LST of the limiting distribution of $(1-\rho)W_i$ $(i = 1, \ldots, N)$ as ρ goes to 1. The expression is strikingly simple and shows explicitly how the waiting-time distributions depend on the system parameters. The results explicitly quantify the trade-off between the increase in fairness and decrease of efficiency introduced by implementing two-phase gated service policies. In particular, the result provide new fundamental insight in the relative waiting-time distributions for one-phase versus two-phase gated service policies. Furthermore, the results reveal a variety of asymptotic insensitivity properties, which provide new insights into the behavior of polling system under medium and heavy load.

The remainder of this paper is organized as follows. In Section 2 the model is described and the main result of the paper is formulated, i.e. a closed-form expression for the LST of the waiting-time distribution in HT, under proper scalings. In Section 3 we discuss a stepwise approach to derive the main result. In Section 4 we discuss several asymptotic properties and address the implications of the results on fairness and efficiency by comparing two-phase polling schemes to the classical one-phase polling model. In Section 5 we address a number of topics for further research. The use of the so-called Descendant Set Approach for the present model is discussed in Appendix A.

2 Model

Consider a system consisting of $N \geq 2$ stations $Q_1, \ldots, Q_N$, each consisting of a phase-1 buffer and a phase-2 buffer. A single server visits and serves the queues in cyclic order. Type-i customers arrive at Q_i according to a Poisson arrival process with rate λ_i, and enter the phase-1 buffer. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i, with Laplace-Stieltjes Transform (LST) $B_i^*(\cdot)$ and with finite k-th moment $b_i^{(k)}$ $(k = 1, 2)$. The k-th moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \sum_{i=1}^{N} \lambda_i b_i^{(k)}/\Lambda$ $(k = 1, 2)$. The load offered to Q_i is $\rho_i = \lambda_i b_i^{(1)}$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i > 0$. Define a polling instant at Q_i as a time epoch at which the server visits Q_i. Each queue is served according to the two-phase gated service policy, which works as follows. When the server arrives at a queue, it closes the gate behind the customers residing in the phase-1 buffer. Then, all customers waiting in the phase-2 buffer are served on a First-Come-First-Served (FCFS) basis. Subsequently, all customers before the gate at the phase-1 buffer are instantaneously forwarded to the phase-2 buffer, and the server proceeds to the next queue. Upon departure from Q_i the server immediately proceeds to Q_{i+1}, incurring a switch-over time R_i, with LST $R_i^*(\cdot)$ and finite k-th moment $r_i^{(k)}$ $(k = 1, 2)$. Denote by $r := \sum_{i=1}^{N} r_i^{(1)} > 0$ the expected total switch-over time per cycle of the server along the queues. All interarrival times, service times and switch-over times are assumed to be mutually independent and independent of the state of the system. A necessary and sufficient condition for the stability of the system is $\rho < 1$ (cf. [9]).

The following notation will be useful. For each variable x that is a function of ρ, we denote its value *evaluated at* $\rho = 1$ by $\hat{x}$. For an event E, denote by I_E the indicator function on E. Moreover, denote by $\mathbf{I}_k$ the k-by-k identity matrix, and by $\mathbf{0}_k$ the k-by-k matrix whose entries are all 0. To make the comparison between the model with two-phase gated defined above to the classical model with one-stage gated service at all queues, we denote by $W_i^{(k)}$ the delay incurred by an arbitrary customer at Q_i, defined as the time between the arrival of a customer at a station and the moment at which it starts to receive service, for the model with k-phase gated service at all queues $(k = 1, 2)$. The main result of this paper is the following closed-form expression for the asymptotic waiting-time distribution for the two-phase gated polling model (defined above) at an arbitrary queue.

Theorem 1 (Main result)
For cyclic polling models with two-phase gated service at each queue, we have for $i = 1, \ldots, N$*:*

$$(1-\rho)W_i^{(2)} \to_d \tilde{W}_i^{(2)} \quad (\rho \uparrow 1) \tag{1}$$

where the LST of $\tilde{W}_i^{(2)}$ *is given by*

$$\tilde{W}_i^{*(2)}(s) = \frac{1}{(1-\hat{\rho}_i)rs}\left\{\left(\frac{\mu_2}{\mu_2 + s(1+\hat{\rho}_i)}\right)^{\alpha_2} - \left(\frac{\mu_2}{\mu_2 + 2s}\right)^{\alpha_2}\right\} \quad (Re(s) > 0), \tag{2}$$

where

$$\alpha_2 := 2r\delta_2 \frac{b^{(1)}}{b^{(2)}}, \quad \mu_2 := 2\delta_2 \frac{b^{(1)}}{b^{(2)}}, \text{ and } \delta_2 := \frac{1}{2}\sum_{j=1}^{N} \hat{\rho}_j(3+\hat{\rho}_j). \tag{3}$$

Here, the limit is taken such that the arrival rates are increased, while keeping both the service-time distributions and the ratios between the arrival rates fixed. The proof of Theorem 1, which is based on a sequence of intermediate results, is given at the end of Section 3. For later reference, we also give a similar result for the case of one-phase gated polling service at all queues (see [33] for a rigorous proof). *For cyclic polling models with one-phase gated service at each queue, we have for* $i = 1, \ldots, N$*:*

$$(1-\rho)W_i^{(1)} \to_d \tilde{W}_i^{(1)} \quad (\rho \uparrow 1) \tag{4}$$

where the LST of $\tilde{W}_i^{(1)}$ *is given by*

$$\tilde{W}_i^{*(1)}(s) = \frac{1}{(1-\hat{\rho}_i)rs}\left\{\left(\frac{\mu_1}{\mu_1 + s\hat{\rho}_i}\right)^{\alpha_1} - \left(\frac{\mu_1}{\mu_1 + s}\right)^{\alpha_1}\right\} \quad (Re(s) > 0), \tag{5}$$

where

$$\alpha_1 := 2r\delta_1 \frac{b^{(1)}}{b^{(2)}}, \quad \mu_1 := 2\delta_1 \frac{b^{(1)}}{b^{(2)}}, \text{ and } \delta_1 := \frac{1}{2}\sum_{j=1}^{N} \hat{\rho}_j(1+\hat{\rho}_j). \tag{6}$$

3 Analysis

In this section we use the theory of MTBPs to derive the main result of the paper, Theorem 1. In Section 3.1 we give a general description of MTBPs, and present a limiting theorem for general MTBPs (Theorem 2) that will be useful throughout. In Section 3.2 we show how the evolution of the polling model under consideration can be described as a MTBP with immigration at each state (Theorem 3). In Section 3.3 we discuss a stepwise approach to combine these results to derive Theorem 1.

3.1 Multi-type branching processes with immigration

In this subsection we give a general description of MTBPs with immigration in each state, and introduce notation useful for further reference. The reader is referred to [3] for more details. We consider a general M-dimensional multi-type branching process with immigration in each state, $\mathbf{Z} = \{\underline{Z}_n, n = 0, 1, \ldots\}$, 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. The process $\mathbf{Z}$ is completely characterized by the one-step offspring function $f(\underline{z}) = (f^{(1)}(\underline{z}), \ldots, f^{(M)}(\underline{z}))$, with $\underline{z} = (z_1, \ldots, z_M)$, and where for $|z_k| \leq 1$ $(k = 1, \ldots, M), i = 1, \ldots, M$,

$$f^{(i)}(\underline{z}) = \sum_{j_1, \ldots, j_M \geq 0} p^{(i)}(j_1, \ldots, j_M) z_1^{j_1} \cdots z_M^{j_M}, \tag{7}$$

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

$$g(\underline{z}) = \sum_{j_1, \ldots, j_M \geq 0} q(j_1, \ldots, j_M) z_1^{j_1} \cdots z_M^{j_M}, \tag{8}$$

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

$$\underline{g} := (g_1, \ldots, g_M), \text{ where } g_i := \frac{\partial g(\underline{z})}{\partial z_i}|_{\underline{z}=\underline{1}} \quad (i = 1, \ldots, M), \tag{9}$$

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

$$\mathbf{M} = (m_{i,j}), \quad \text{with } m_{i,j} := \frac{\partial f^{(i)}(\underline{z})}{\partial z_j}|_{\underline{z}=\underline{1}} \quad (i, j = 1, \ldots, M). \tag{10}$$

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

$$\mathbf{K}^{(i)} = \left(k_{j,k}^{(i)}\right), \quad \text{with } k_{j,k}^{(i)} := \frac{\partial^2 f^{(i)}(\underline{z})}{\partial z_j \partial z_k}|_{\underline{z}=\underline{1}} \quad (i, j, k = 1, \ldots, M). \tag{11}$$

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, or the Perron-Frobenius eigenvalue (cf., e.g., [3]), normalized such that

$$\underline{v}^\top \underline{1} = \underline{v}^\top \underline{w} = 1. \tag{12}$$

The following conditions are necessary and sufficient conditions for the ergodicity of the process $\mathbf{Z}$ (cf. [24, 28]): $\xi < 1$ and

$$\sum_{j_1+\cdots+j_M>0} q(j_1, \ldots, j_M)\log(j_1 + \cdots + j_M) < \infty. \tag{13}$$

Following standard branching-process terminology the process $\mathbf{Z}$ is called sub-critical if $\xi < 1$, critical if $\xi = 1$ and super-critical if $\xi > 1$. 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 $\xi > 0$ let

$$\pi_0(\xi) := 0, \quad \text{and} \quad \pi_n(\xi) := \sum_{r=1}^{n} \xi^{r-2}, \quad n = 1, 2, \ldots. \tag{14}$$

Definition
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

$$f_\Gamma(x) = \frac{\mu^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\mu x} \quad (x > 0) \quad \text{with} \quad \Gamma(\alpha) := \int_{t=0}^{\infty} t^{\alpha-1} e^{-t} dt, \tag{15}$$

and Laplace-Stieltjes Transform (LST)

$$\Gamma^*(s) = \left(\frac{\mu}{\mu + s}\right)^\alpha \quad (Re(s) > 0). \tag{16}$$

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

Theorem 2
Assume that all derivatives of $f(\underline{z})$ of order two exist at $\underline{z} = \underline{1}$ and that $0 < g_i < \infty$ $(i = 1, \ldots, M)$. Then

$$\frac{1}{\pi_n(\xi)} \begin{pmatrix} Z_n^{(1)} \\ \vdots \\ Z_n^{(M)} \end{pmatrix} \to_d A \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_M \end{pmatrix} \Gamma(\alpha, 1) \quad as \quad (\xi, n) \to (1, \infty), \tag{17}$$

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

$$\alpha := \frac{1}{A}\underline{\hat{g}}^\top \underline{\hat{w}} = \frac{1}{A}\sum_{i=1}^{M} \hat{g}_i \hat{w}_i, \ \textit{with} \ \ A := \sum_{i=1}^{M} \hat{v}_i \left(\underline{\hat{w}}^\top \hat{\mathbf{K}}^{(i)} \underline{\hat{w}}\right) > 0. \tag{18}$$

Proof: See [24]. We refer to Remark 3.3 for the details about the convergence and the limiting regime considered in (17). □

In the next two subsections we will show how Theorem 2, which was derived in the context of generic MTBPs, can be transformed into results for the two-phase gated polling model under consideration.

3.2 Preliminaries

Without loss of generality, throughout we will focus on the waiting times at Q_1 and consider the state of the system at polling instants at Q_1. Let $X_i^{(k)}$ be the number of phase-k customers at Q_i at an arbitrary polling instant at Q_1 when the system is in steady state ($k = 1, 2,\ i = 1, \ldots, N$). Moreover, for $i = 1, \ldots, N$, we consider the two-dimensional random variable $\left(X_i^{(1)}, X_i^{(2)}\right)$, and denote the corresponding Probability Generating Function (PGF) by, for $|z_1|, |z_2| \leq 1$,

$$X_i^*(z_1, z_2) := E\left[z_1^{X_i^{(1)}} z_2^{X_i^{(2)}}\right]. \tag{19}$$

Denoting the LST of the waiting-time distribution at Q_1 by $W_1^{*(2)}(\cdot)$, the waiting-time distribution at Q_1 is related to the joint distribution of $\left(X_1^{(1)}, X_1^{(2)}\right)$ through the following expression (cf. [20]): For $Re\ s \geq 0,\ \rho < 1$,

$$W_1^{*(2)}(s) = \frac{X_1^*(1 - s/\lambda_1, B_1^*(s)) - X_1^*(1 - s/\lambda_1, 1 - s/\lambda_1)}{E\left[X_1^{(1)}\right](B_1^*(s) - 1 + s/\lambda_1)}. \tag{20}$$

To start the analysis, note first that straightforward balancing arguments lead to the following expression for the first moments $E\left[X_1^{(k)}\right]$ $(k = 1, 2)$: For $\rho < 1$,

$$E\left[X_1^{(1)}\right] = E\left[X_1^{(2)}\right] = \frac{\lambda_1 r}{1 - \rho}. \tag{21}$$

In general the distributions and moments of $X_1^{(1)}$ and $X_1^{(2)}$ can not be obtained explicitly. The following notation is useful. Let

$$\underline{X} := \left(X_1^{(1)}, \ldots, X_N^{(1)}, X_1^{(2)}, \ldots, X_N^{(2)}\right) \tag{22}$$

be the $2N$-dimensional vector that describes the state of the system at an arbitrary polling instant at Q_1. To determine the asymptotic behavior of the waiting-time distribution given in (20), we focus on the limiting behavior of $\underline{X}$ as ρ goes to 1. To this end, below we describe the evolution of the system as an MTBP. Subsequently, in Section 3.3 we use this to transform Theorem 2 into an asymptotic expression for the distribution of $(1-\rho)\underline{X}$, i.e. the scaled version of $\underline{X}$ as ρ goes to 1.

To establish the relation with the general MTBP-model described in Section 2, let $X_{i,n}^{(k)}$ be the number of type-i customers at phase-k in the system at the n-th polling instant at Q_1, for $i = 1, \ldots, N$, $k = 1, 2$ and $n = 0, 1, \ldots$, and let

$$\underline{X}_n := \left(X_{1,n}^{(1)}, \ldots, X_{N,n}^{(1)}, X_{1,n}^{(2)}, \ldots, X_{N,n}^{(2)}\right) \tag{23}$$

be the state vector at the n-th polling instant at Q_1. Then similar to the analysis made by Resing [28] we make the following observation.

Theorem 3
The discrete-time process $\{\underline{X}_n, n = 0, 1, \ldots\}$ constitutes a $2N$-dimensional MTBP with immigration in each state, the LST of the offspring function is given by the following expression: For $|s_i^{(k)}| \leq 1$ $(i = 1, \ldots, N, k = 1, 2)$,

$$f(\underline{s}) := \left(f^{(1)}(\underline{s}), \ldots, f^{(2N)}(\underline{s})\right),\ \textit{with}\ \underline{s} := \left(s_1^{(1)}, \ldots, s_N^{(1)}, s_1^{(2)}, \ldots, s_N^{(2)}\right), \tag{24}$$

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

$$f^{(i)}\left(s_1^{(1)}, \ldots, s_N^{(1)}, s_1^{(2)}, \ldots, s_N^{(2)}\right) := s_i^{(2)}, \tag{25}$$

and

$$f^{(i+N)}\left(s_1^{(1)},\ldots,s_N^{(1)},s_1^{(2)},\ldots,s_N^{(2)}\right) := B_i^*\left(\sum_{j=1}^{i}\lambda_j\left(1-s_j^{(1)}\right)+\sum_{j=i+1}^{N}\lambda_j\left(1-s_j^{(2)}\right)\right), \quad (26)$$

and where the LST of the immigration function is given by

$$g\left(s_1^{(1)},\ldots,s_N^{(1)},s_1^{(2)},\ldots,s_N^{(2)}\right) := \prod_{i=1}^{N} R_i^*\left(\sum_{j=1}^{i}\lambda_j\left(1-s_j^{(1))}\right)+\sum_{j=i+1}^{N}\lambda_j\left(1-s_j^{(2)}\right)\right). \quad (27)$$

Proof: Relations (25)-(27) can be obtained along the lines of [28] for the case of one-phase gated service, using simple generating-function manipulations. More specifically, in the spirit of the work in [28], equation (25) follows from the fact that a type-i customer at phase-1 at a given polling instant P_1 at Q_1 is "effectively replaced" by a single type-i customer at phase-2 at the next polling instant at Q_1; note that in this case, the type-i customer is simply forwarded from phase-1 to phase-2. Similarly, (26) follows from the fact that a type-i customer at phase-2 at P_1 is effectively replaced by all customers that arrive in the system during its service time with LST $B_i^*(\cdot)$. Finally, (27) stems from the fact that the immigration consists of the contributions of newly arriving customers that arrive during the switch-over times, which are independently distributed with LST $R_i^*(\cdot), i = 1,\ldots,N$. □

3.3 Stepwise derivation of Theorem 1

In this section we use Theorem 3 to transform Theorem 2 into an expression for the limiting distribution for $(1-\rho)\underline{X}$ as ρ goes to 1 (Theorem 4). Subsequently, this result is combined with (20) to prove Theorem 1. The stepwise approach consists of the following steps:

Step 1: Derive an expression for the mean offspring matrix $\mathbf{M}$ for the polling model under consideration (Lemma 1).
Step 2: Derive an expression for the left and right eigenvectors $\underline{v}$ and $\underline{w}$ of the mean matrix, evaluated at $\rho = 1$ (Lemma 2)
Step 3: Derive an expression for the mean immigration vector $\underline{g}$, evaluated at $\rho = 1$ (Lemma 3).
Step 4: Derive an expression for limiting behavior of $\xi(\rho)$ considered as a function of ρ, as ρ goes to 1 (Lemma 4).
Step 5: Derive an expression for A, evaluated at $\rho = 1$ (Lemma 5).
Step 6: Combine steps 1 to 5 into an asymptotic expression for the distribution of $(1-\rho)\underline{X}$ as ρ goes to 1.
Step 7: Use this expression in combination with (20) to obtain Theorem 1.

In Subsections 3.3.1 to 3.3.7 each of the steps will be discussed in more detail below.

3.3.1 Step 1: Mean matrix

The following result gives an expression for the offspring matrix for the polling model under consideration.

Lemma 1
For the two-phase gated polling model, the mean offspring matrix $\mathbf{M} = (m_{i,j})$ *is given by*

$$\mathbf{M} = \mathbf{M_1}\cdots\mathbf{M}_N\mathbf{P}, \quad (28)$$

where $\mathbf{P}$ *is the permutation block matrix*

$$\mathbf{P} = \begin{pmatrix} \mathbf{0}_N & \mathbf{I}_N \\ \mathbf{I}_N & \mathbf{0}_N \end{pmatrix}, \quad (29)$$

where for $k = 1, \ldots, N$, *the elements of the matrix* $\mathbf{M}_k = \left(m_{i,j}^{(k)}\right)$ *are given by: For* $i, j = 1, \ldots, 2N$, $i \neq N + k$,

$$m_{i,j}^{(k)} = I_{\{i=j\}}, \tag{30}$$

and for $i = N + k$,

$$m_{N+k,k+j}^{(k)} = \lambda_{j+k} b_k^{(1)} \text{ for } j = 1, \ldots, N - k, \tag{31}$$

$$m_{N+k,k+j}^{(k)} = \lambda_{j+k-N} b_k^{(1)} \text{ for } j = N - k + 1, \ldots, N, \tag{32}$$

$$m_{N+k,j}^{(k)} = 0 \text{ for } j = 1, \ldots, k \text{ or } j = N + k + 1, \ldots, 2N. \tag{33}$$

Proof: The result can be obtained in a tedious but fairly straightforward manner by taking the partial derivatives of the offspring function defined in (25) and (26). As an alternative, we can use the description of the Descendant Set Approach (DSA) discussed in Appendix A, and define for $c = -1, 0, 1, \ldots$ the following vector of descendant set variables, defined in (83)-(85):

$$\underline{\alpha}_c := \left(\alpha_{1,c}^{(1)} \ \cdots \ \alpha_{N,c}^{(1)} \ \alpha_{1,c}^{(2)} \ \cdots \ \alpha_{N,c}^{(2)}\right)^\top = \left(\alpha_{1,c-1}^{(2)} \ \cdots \ \alpha_{N,c-1}^{(2)} \ \alpha_{1,c}^{(2)} \ \cdots \ \alpha_{N,c}^{(2)}\right)^\top, \tag{34}$$

Then from DSA point-of-view it is readily seen that it suffices to show that the matrix $\mathbf{M}$ defined in (28)-(33) satisfies the one-step relation $\underline{\alpha}_{c+1} = \mathbf{M}\underline{\alpha}_c$, for $c = -1, 0, 1, \ldots$. To this end, we first note that using relation (34) it is easily verified that, for $c = -1, 0, 1, \ldots$

$$\mathbf{P}\underline{\alpha}_c = \left(\alpha_{1,c}^{(2)} \ \cdots \ \alpha_{N,c}^{(2)} \ \alpha_{1,c-1}^{(2)} \ \cdots \ \alpha_{N,c-1}^{(2)}\right)^\top. \tag{35}$$

Then using equations (84)-(86), it is readily verified (by induction in k, for $k = N, N-1, \ldots, 1$) that for $k = 1, \ldots, N$, $c = -1, 0, 1, \ldots$,

$$\mathbf{M}_k \ldots \mathbf{M}_N \mathbf{P}\underline{\alpha}_c = \left(\alpha_{1,c}^{(2)} \ \ldots \ \alpha_{N,c}^{(2)} \ \alpha_{1,c-1}^{(2)} \ \cdots \ \alpha_{k-1,c-1}^{(2)} \ \alpha_{k,c+1}^{(2)} \ \cdots \ \alpha_{N,c+1}^{(2)}\right)^\top, \tag{36}$$

which immediately implies by taking $k = 1$ that

$$\mathbf{M}\underline{\alpha}_c = \mathbf{M}_1 \ldots \mathbf{M}_N \mathbf{P}\underline{\alpha}_c = \left(\alpha_{1,c}^{(2)} \ \cdots \ \alpha_{N,c}^{(2)} \ \alpha_{1,c+1}^{(2)} \ \cdots \ \alpha_{N,c+1}^{(2)}\right)^\top = \underline{\alpha}_{c+1}. \ \square \tag{37}$$

3.3.2 Step 2: Left and right eigenvectors of M at $\rho = 1$

The following result gives the left and right eigenvectors of the mean offspring matrix $\mathbf{M}$ defined in Lemma 1 above, evaluated at $\rho = 1$.

Lemma 2
For the two-phase gated polling model, the right and left eigenvectors of the mean matrix $\hat{\mathbf{M}}$ *are given by*

$$\underline{\hat{w}} = \begin{pmatrix} \hat{w}_1 \\ \vdots \\ \hat{w}_{2N} \end{pmatrix} := |\underline{b}|^{-1}\underline{b}, \ \textit{with} \ \underline{b} := \begin{pmatrix} b_1^{(1)} \\ \vdots \\ b_N^{(1)} \\ b_1^{(1)} \\ \vdots \\ b_N^{(1)} \end{pmatrix}, \tag{38}$$

and

$$\underline{\hat{v}} = \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_{2N} \end{pmatrix} := \frac{|\underline{b}|}{\delta}\underline{\hat{u}}, \ \textit{with} \ \underline{u} := \begin{pmatrix} u_1 \\ \vdots \\ u_{2N} \end{pmatrix}, \quad \textit{where} \ u_i := \lambda_i(\rho_i + \cdots + \rho_N), \quad u_{i+N} := \lambda_i \quad (i = 1, \ldots, N), \tag{39}$$

and where

$$\delta := \underline{\hat{u}}^\top \underline{b} = \sum_{i=1}^{N}\sum_{j=i}^{N} \hat{\rho}_i\hat{\rho}_j + \sum_{i=1}^{N}\hat{\rho}_i = \frac{1}{2}\left(1 + \sum_{i=1}^{N}\hat{\rho}_i^2\right) + 1 = \frac{1}{2}\sum_{i=1}^{N}\hat{\rho}_i(3+\hat{\rho}_i), \ \textit{and} \ |\underline{b}| := 2\sum_{k=1}^{N} b_k^{(1)}. \tag{40}$$

Proof: First, it is readily seen by using equations (30)-(33) that for $k = 1, \ldots, N$, we have

$$\sum_{j=1}^{2N} m_{N+k,j}^{(k)} \hat{w}_j = |\underline{b}|^{-1}\left(\sum_{j=1}^{N-k} \hat{\lambda}_{j+k} b_k^{(1)} b_j^{(1)} + \sum_{j=N-k+1}^{N} \hat{\lambda}_{j+k-N} b_k^{(1)} b_j^{(1)}\right) = |\underline{b}|^{-1} b_k^{(1)} \sum_{j=1}^{N} \hat{\lambda}_j b_j^{(1)} = |\underline{b}|^{-1} b_k^{(1)} = \hat{w}_{N+k}. \tag{41}$$

This immediately implies that $\hat{\mathbf{M}}_k \underline{\hat{w}} = \underline{\hat{w}}$ for $k = 1, \ldots, N$. Moreover, it is easy to see that $\mathbf{P}\underline{\hat{w}} = \underline{\hat{w}}$. Combining these observations then implies $\mathbf{M}\underline{\hat{w}} = \mathbf{M}_1 \cdots \mathbf{M}_N \mathbf{P}\underline{\hat{w}} = \underline{\hat{w}}$, which shows that $\underline{\hat{w}}$ indeed is a right eigenvector of $\hat{\mathbf{M}}$. Similar arguments can be used to show that $\hat{\mathbf{M}}^\top \underline{\hat{v}} = \underline{\hat{v}}$. The details of the proof are omitted for compactness of the presentation. □

3.3.3 Step 3: Mean immigration vector $\underline{g}$

We now proceed to specify the mean immigration vector $\underline{g}$, defined in (9), for the model under consideration. Considering the evolution of the $2N$-dimensional state vector as a discrete-time Markov chain $\{\underline{X}_n, n = 0, 1, \ldots\}$ at successive polling instants at Q_1, the "immigrants" in the n-th generation are the customers present a time n that are not children of any of the customers present at time $n-1$. Denote the mean immigration vector by

$$\underline{g} = \left(g_1^{(1)} \ \cdots \ g_N^{(1)} \ g_1^{(2)} \ \cdots \ g_N^{(2)}\right)^\top, \tag{42}$$

where $g_i^{(k)}$ stands for the mean number of type-i immigrants in phase-k. In the descendant set point-of-view, $g_i^{(k)}$ can be seen as the mean number of type-i customers in phase-k present at the reference point (at Q_1) that arrived during a switch-over time in cycle 0 (i.e. the time between the reference point and the preceding polling instant at Q_1).

Lemma 3
For the two-phase gated model, for $j = 1, \ldots, N$,

$$g_j^{(1)} = \lambda_j \sum_{i=j}^{N} r_i^{(1)}, \tag{43}$$

$$g_j^{(2)} = \lambda_j \sum_{i=1}^{j-1} r_i^{(1)} \tag{44}$$

and

$$\underline{\hat{g}}^\top \underline{\hat{w}} = |\underline{b}|^{-1} r. \tag{45}$$

Proof: Equations (43) and (44) can be directly obtained from (27). To give a more intuitive derivation, note that equation (43) follows directly from the fact that a type-j immigrant in phase-1 should have arrived during a switch-over time from Q_i to Q_{i+1}, for some $i \geq j$. Similarly, equation (44) follows directly from the fact that a type-j immigrant in phase-2 should have arrived during a switch-over time from Q_i to Q_{i+1}, for some $i < j$. Finally, to prove (45), assume $\rho = 1$. Then (45) follows directly from the following sequence of relations:

$$\hat{\underline{g}}^{\top} \hat{\underline{w}} := \sum_{j=1}^{2N} \hat{g}_j \hat{w}_j = \sum_{j=1}^{N} \hat{g}_j^{(1)} \hat{w}_j + \sum_{j=1}^{N} \hat{g}_j^{(2)} \hat{w}_{N+j} = |\underline{b}|^{-1} \left(\sum_{j=1}^{N} b_j^{(1)} \hat{\lambda}_j \sum_{i=j}^{N} r_i^{(1)} + \sum_{j=1}^{N} b_j^{(1)} \hat{\lambda}_j \sum_{i=1}^{j-1} r_i^{(1)} \right) \tag{46}$$

$$= |\underline{b}|^{-1} \left(\sum_{j=1}^{N} b_j^{(1)} \hat{\lambda}_j r \right) = |\underline{b}|^{-1} \hat{\rho} r = |\underline{b}|^{-1} r. \ \square \tag{47}$$

3.3.4 Step 4: Limiting behavior of $\xi(\rho)$ for $\rho \uparrow 1$

The following result describes the limiting behavior of the maximum eigenvalue $\xi(\rho)$ of the matrix $\mathbf{M}$ defined in Lemma 1, considered as a function of ρ, as ρ goes to 1.

Lemma 4
For the two-phase gated polling model, the maximum eigenvalue $\xi = \xi(\rho)$ satisfies 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*

$$\xi'(1) = \lim_{\rho \uparrow 1} \frac{1 - \xi(\rho)}{1 - \rho} = \frac{1}{\delta}, \tag{48}$$

where δ *is defined in* (40).

Proof: Part 1 was shown in [28]. Part 2 follows from the fact that all entries of $\mathbf{M}$ are continuous functions of ρ, which implies the continuity of $\xi(\rho)$ with respect to ρ (see for example [3]). 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 model under consideration is work conserving. Finally, to prove Part 4 we adopt the concept of the Descendant Set Approach (DSA) discussed in Appendix A. Then, based on known properties for the maximum eigenvalue of positive semi-definite matrices applied to $\mathbf{M}$ (see for example [3]) we can decompose $\alpha_{i,c}^{(1)}$, defined in (83)-(85), into a dominant and a recessive part as follows: For $\rho < 1$, $i = 1, \ldots, N$,

$$\alpha_{i,c}^{(1)} = \xi^{c+1} w_i v_1 + s_{i,c}^{(1)} \tag{49}$$

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

$$\sum_{c=0}^{\infty} s_{i,c}^{(1)} < \infty. \tag{50}$$

The result then follows directly from (49), (50), (21), (87) and Parts 1, 2 and 3 of Lemma 4. This completes the proof. $\square$

3.3.5 Step 5: Expression for A at $\rho = 1$

Lemma 5
For the two-phase gated polling model,

$$A = |\underline{b}|^{-1}\delta^{-1}\frac{b^{(2)}}{2b^{(1)}}. \tag{51}$$

Proof: Using the definition of A in Theorem 2, we need to specify the eigenvectors $\hat{v}$ and $\hat{w}$ of the mean offspring matrix $\hat{\mathbf{M}}$, and the second-order matrices $\hat{\mathbf{K}}^{(i)}$ $(i = 1, \ldots, N)$. To this end, note that $\hat{v}$ and $\hat{w}$ are given in Lemma 2. The matrices $\hat{\mathbf{K}}^{(i)}$ $(i = 1, \ldots, N)$ can be obtained directly from Theorem 3. This method is methodologically straightforward, but practically quite cumbersome; the details of this derivation are left as an exercise to the reader. As an alternative, the scaling constant A in (51) can also be obtained by simple first-order arguments only, see Remark 3.1 below. □

3.3.6 Step 6: Asymptotic expression for scaled state vector

We are now ready to present the HT result for the state vector at polling instants in the two-phase gated polling model. Without loss of generality, we focus on the evolution of the state vector at embedded polling instants at Q_1.

Theorem 4
For the two-phase gated polling model, the state vector at polling instants at Q_1 has the following asymptotic behavior:

$$(1-\rho)\underline{X}^{\top} \to_d \delta \cdot A \cdot \underline{\hat{v}} \cdot \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{52}$$

where

$$\alpha = 2r\delta\frac{b^{(1)}}{b^{(2)}}. \tag{53}$$

and where δ, A and $\hat{v}_i$ $(i = 1, \ldots, 2N)$ are defined in (40), (51) and (39), respectively.

Proof: To start, note that the process that describes the evolution of the state vector $\{\underline{X}_n, n = 0, 1, \ldots\}$ at successive polling instants at Q_1 constitutes an $2N$-dimensional MTBP with offspring function $f(\underline{z})$ and immigration function $g(\underline{z})$ defined in Theorem 3, and with mean matrix $\mathbf{M}$ defined in Lemma 1. Moreover, from Theorem 3 and Lemma 3 it is readily verified that the assumptions of Theorem 2 on the finiteness of the second-order derivatives of $f(\underline{s})$ and the mean immigration function $\underline{g}$ are satisfied (with $M = 2N$), based on the assumption that the second moments of the service times $b_i^{(2)}$ $(i = 1, \ldots, N)$ and the first moments of the switch-over times $r_i^{(1)}$ $(i = 1, \ldots, N)$ are finite. Then using Lemmas 2 to 4 and Theorem 2 it follows that

$$\frac{1}{\pi_n(\xi(\rho))} \cdot \underline{X}_n^{\top} \to_d A \cdot \underline{\hat{v}} \cdot \Gamma(\alpha, 1) \quad as \quad (\rho, n) \to (1, \infty), \tag{54}$$

where A, $\underline{\hat{v}}$ and α are defined in (17)-(18), see Remark 3.3 below for more details about the convergence and the limiting regime. Hence, using the properties listed in Lemma 4, it readily follows from (54) that

$$(1-\rho)\underline{X}_n^{\top} \to_d \delta \cdot A \cdot \underline{\hat{v}} \cdot \Gamma(\alpha, 1) \quad as \quad (\rho, n) \to (1, \infty). \tag{55}$$

□

Remark 3.1
As an alternative to the proof of Lemma 5, a much simpler derivation of A can be obtained by

combining Theorem 4 (which does not require the scaling parameter A to be known explicitly), and (21). To this end, note that by taking the mean value of the first entry in (52) and taking the limit for $\rho \uparrow 1$ it readily follows that

$$\hat{\lambda}_1 r = \delta A \hat{v}_1 \cdot \alpha = \delta A \cdot \frac{|\underline{b}|}{\delta} \hat{\lambda}_1 \cdot 2r\delta \frac{b^{(1)}}{b^{(2)}}, \tag{56}$$

which immediately implies (51). $\square$

Remark 3.2
The parameters δ and $\hat{v}$ in (52) only depend on the arrival rates and the *mean* values of the service-time distributions. Therefore, the impact of the variability of the service-time distributions manifests itself in the parameters α (defined in (53)) and A (given in Lemma 5). From Lemma 5 it follows that A is the mean residual service time of an *arbitrary* customer (regardless of the queue it enters), up to a normalizing constant. This normalizing constant $|\underline{b}|^{-1}\delta^{-1}$ could be set to one by properly renormalizing the eigenvectors of the mean matrix $\mathbf{M}$ in Lemma 2.

Remark 3.3
A note on the convergence in Theorems 2 and 4. The convergence of Theorem 2 should be considered in the following sense (see [24] for details): for all $\epsilon > 0$ there exist $\delta > 0$ and N such that if $|1 - \xi| < \delta$ then for all $n > N$ it holds that

$$\sup_{\underline{x} \in \mathcal{R}^M} \left| \text{Prob}\left\{ \frac{1}{\pi_n(\xi)} \underline{Z}_n \leq \underline{x} \right\} - \text{Prob}\left\{ A \cdot \Gamma(\alpha, 1) \cdot \underline{\hat{v}} \leq \underline{x} \right\} \right| < \epsilon, \tag{57}$$

where ξ, $\pi_n(\cdot)$, $\underline{Z}_n$, A, α and $\underline{v}$ are defined in Section 3.1. And similarly, for the polling model under consideration the convergence in (52) is defined as follows: for all $\epsilon > 0$ there exist $\delta > 0$ and N such that if $|1 - \rho| < \delta$ then for all $n > N$ it holds that

$$\sup_{\underline{x} \in \mathcal{R}^{2N}} \left| \text{Prob}\left\{ (1-\rho)\underline{X}_n^{\top} \leq \underline{x} \right\} - \text{Prob}\left\{ A \cdot \Gamma(\alpha, 1) \cdot \underline{\hat{v}} \leq \underline{x} \right\} \right| < \epsilon, \tag{58}$$

where $\underline{X}_n$, A, α and $\underline{v}$ are specified for the two-phase gated polling model in Section 3.3. Using these definitions, it is easily seen that Theorem 2 translates into Theorem 4 by using relations (54)-(55), using Theorem 3 and the properties listed in Lemma 4.

We are now ready to formulate the proof of Theorem 1.

3.3.7 Proof of Theorem 1

Without loss of generality, we assume $i = 1$. Throughout it will be convenient to relate the waiting-time and queue-length distributions at polling instants at Q_1 to the joint distribution of *two successive cycle times*. More precisely, let a given polling instant P at Q_1 mark the end of a cycle time with duration C_1, and let the duration of the *preceding* cycle time be C_1^-. Moreover, denote the joint LST of (C_1, C_1^-) by: For $Re(s), Re(t) > 0$,

$$C_1^*(s,t) := E\left[e^{-sC_1 - tC_1^-} \right]. \tag{59}$$

Recall from (19) that $X_1^*(z_1, z_2)$ is the joint PGF of the numbers of type-1 customers at both phases at queue 1 at an arbitrary polling instant at Q_1. Then the population of customers present at Q_1 at phase 1 at polling instant P consists exactly of those customers that arrived during the past cycle time of duration C_1, whereas the population of customers present at phase 2 exactly consists of those that arrived in the preceding cycle of duration C_1^-. Standard GF manipulations then immediately imply that for $|z_1|, |z_2| \leq 1$,

$$X_1^*(z_1, z_2) = C_1^*\left(\lambda_1(1 - z_1), \lambda_1(1 - z_2)\right). \tag{60}$$

Using (60), equation (20) can be reformulated in the following convenient form: For $Re(s) > 0$,

$$W_1^*(s) = \frac{(1-\rho_1)s}{s - \lambda_1(1 - B_1^*(s))} \cdot \frac{C_1^*(s, \lambda_1(1 - B_1^*(s))) - C_1^*(s,s)}{s(1-\rho_1)r/(1-\rho)}. \tag{61}$$

Now, combining (39) with Lemma 5 and Theorem 3, by taking the first and the $(N+1)$-st component of the vector in (52), it readily follows that

$$(1-\rho) \begin{pmatrix} X_1^{(1)} \\ X_1^{(2)} \end{pmatrix} \to_d \frac{1}{2\delta} \cdot \frac{b^{(2)}}{b^{(1)}} \begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_1 \end{pmatrix} \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{62}$$

where α is defined in (53). Then, using (60) and similar arguments as those discussed in [33], equation (62) can be expressed in terms of cycle times as

$$(1-\rho) \begin{pmatrix} C_1 \\ C_1^- \end{pmatrix} \to_d \frac{1}{2\delta} \cdot \frac{b^{(2)}}{b^{(1)}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{63}$$

where convergence should be considered in the supremum-sense defined in (58). Theorem 1 follows then directly by combining (61), (63) and (16) and standard algebraic manipulations, recalling that we focused on the waiting-time distributions Q_1 without loss of generality. $\square$

4 Discussion and Implications

Theorem 1 reveals a variety of asymptotic properties of the performance of the model under consideration. In Section 4.1 we formulate several insensitivity properties of the asymptotic waiting-time distributions with respect to the system parameters. In Section 4.2 we assess the implications regarding the trade-off between efficiency and fairness, comparing the performance of two-phase gated model to the classical one-phase gated model.

4.1 Insensitivity

The following result follows directly from Theorem 1.

Property 1 (Insensitivity)
For $i = 1, \ldots, N$, the distribution of $\tilde{W}_i^{(2)}$

(1) *is independent of the visit order,*
(2) *depends on the switch-over time distributions only through r, i.e., the total expected switch-over time per cycle,*
(3) *depends on the higher moments of the service-time distributions only through $b^{(2)}$, i.e., the second moment of the service time of an arbitrary customer.*

In general, Proposition 1 is not valid for stable systems (i.e., for $\rho < 1$), where the visit order, the complete service-time and switch-over time distributions *do* have an impact on the waiting-times distributions. Hence, Proposition 1 shows that the influence of these parameters on the waiting-time distributions vanishes when the load tends to unity, and as such can be viewed as lower-order effects in heavy traffic. Note that it follows from (4)-(6) that same insensitivity properties are valid for the distribution of $\tilde{W}_i^{(1)}$, defined in (4)-(5), see also [33].

4.2 Efficiency and fairness

Theorem 1 also leads to several interesting insights regarding the trade-off between efficiency and fairness when comparing models with one-phase and two-phase gated service.

4.2.1 Efficiency

A commonly used measure of efficiency is the mean total amount of unfinished work in the system. To this end, let $V^{(k)}$ denote the total amount of waiting work in the system, for the model with k-phase gated service ($k = 1, 2$). Note that it follows directly from Little's formula that, for $\rho < 1$, $E[V^{(k)}] = \sum_{i=1}^{N} \rho_i E[W_i^{(k)}]$. Then the following results follows directly from the pseudo-conservation law (cf. [39] for details of the derivation).

Property 2 (Efficiency)
Two-phase gated service is less efficient than one-phase gated service in the sense that, for $\rho < 1$,

$$E[V^{(2)}] = E[V^{(1)}] + \frac{r}{1-\rho} > E[V^{(1)}]. \tag{64}$$

Note that Property 2 immediately implies that the one-phase gated model is also *asymptotically* more efficient than the two-phase gated model, in the sense that if we define $v^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)E[V^{(k)}]$ ($k = 1, 2$), then $v^{(2)} = v^{(1)} + r > v^{(1)}$. We refer to [39] for numerical results that support these observations.

4.2.2 Fairness

While efficiency in polling models has been extensively studied and is quite well understood, there is no commonly agreed upon theoretical yardstick for measuring fairness in polling models. One can think of different notions of fairness in polling systems. In this section we consider two types of fairness: (1) *queue fairness* and (2) *customer fairness*. Intuitively, ultimate queue fairness is realized when the waiting-time distributions are the same for all queues, while ultimate customer fairness is reached when the customers somehow experience the same *slowdown*. For completeness, we first outline the results on queue fairness (as discussed in more detail in equations (24), (25) and Table 1 in [39]). Subsequently, we discuss the implications of Theorem 1 on customer fairness.

Queue fairness
Let us define the *queue unfairness* as follows (see also [39]): For $\rho < 1$,

$$\mathcal{F}_{queue}^{(k)} := \max_{i,j=1,\ldots,N} \left| \frac{E\left[W_i^{(k)}\right]}{E\left[W_j^{(k)}\right]} - 1 \right| \quad (k = 1, 2). \tag{65}$$

Thus, the higher the (asymptotic) unfairness, the *less* fair is the service policy. Now, it follows directly from (4)-(6) above that, for $i = 1, \ldots, N$,

$$E\left[\tilde{W}_i^{(1)}\right] = \frac{1+\hat{\rho}_i}{4\delta_1} \frac{b^{(2)}}{b^{(1)}} + \frac{r(1+\hat{\rho}_i)}{2}, \quad \text{with } \delta_1 := \frac{1}{2} \sum_{j=1}^{N} \hat{\rho}_j (1+\hat{\rho}_j), \tag{66}$$

and from Theorem 1 that, for $i = 1, \ldots, N$,

$$E\left[\tilde{W}_i^{(2)}\right] = \frac{3+\hat{\rho}_i}{4\delta_2} \frac{b^{(2)}}{b^{(1)}} + \frac{r(3+\hat{\rho}_i)}{2}, \quad \text{with } \delta_2 := \frac{1}{2} \sum_{j=1}^{N} \hat{\rho}_j (3+\hat{\rho}_j). \tag{67}$$

The following result is an immediate consequence of (66) and (67).

Property 3 (Asymptotic queue fairness)
Two-phase gated service is asymptotically more fair than one-phase gated service in the sense that

$$\hat{\mathcal{F}}_{queue}^{(2)} < \hat{\mathcal{F}}_{queue}^{(1)}. \tag{68}$$

This observation follows directly from the fact that *for all* $i, j = 1, \ldots, N$. Then by taking the limit for $\rho \uparrow 1$ it holds that

$$\left| \frac{E\left[\tilde{W}_i^{(2)}\right]}{E\left[\tilde{W}_j^{(2)}\right]} - 1 \right| = \left| \frac{3+\hat{\rho}_i}{3+\hat{\rho}_j} - 1 \right| < \left| \frac{1+\hat{\rho}_i}{1+\hat{\rho}_j} - 1 \right| = \left| \frac{E\left[\tilde{W}_i^{(1)}\right]}{E\left[\tilde{W}_j^{(1)}\right]} - 1 \right|, \tag{69}$$

which is easily seen to imply (68), by taking the limit for $\rho \uparrow 1$. We refer to [39] for numerical results that support these observations.

Customer fairness

Let $W^{(k)}$ be the waiting time of an *arbitrary* customer in the k-phase gated polling system, and recall that the m-th moment of its service time is denoted by $b^{(m)} = \Lambda^{-1} \sum_{i=1}^{N} \lambda_i b_i^{(m)}$ $(m = 1, 2)$, for both the one-phase and two-phase gated model. Then Brosh et al. [6] define the following Slowdown Queueing Fairness (SQF) measure of customer unfairness for the polling model with k-phase gated service at all queues: For $k = 1, 2$,

$$\mathcal{F}_{SQF}^{(k)} := Var\left[W^{(k)} - \frac{E\left[W^{(k)}\right]}{b^{(1)}} B \right] = E\left[W^{(k)} - \frac{E\left[W^{(k)}\right]}{b^{(1)}} B \right]^2. \tag{70}$$

Then the asymptotic results presented in Theorem 1 can be used to quantify $\mathcal{F}_{SQF}^{(k)}$ in the limiting case $\rho \uparrow 1$. More precisely, the asymptotic fairness for the polling model is given by the following expression: For $k = 1, 2$,

$$\hat{\mathcal{F}}_{SQF}^{(k)} := E\left[\tilde{W}^{(k)} - \frac{E\left[\tilde{W}^{(k)}\right]}{b^{(1)}} B \right]^2 = E\left[\left(\tilde{W}^{(k)}\right)^2 \right] - 2E\left[\tilde{W}^{(k)}\right] \sum_{i=1}^{N} \hat{\rho}_i E\left[\tilde{W}_i^{(k)}\right] + \left(E\left[\tilde{W}^{(k)}\right]\right)^2 \cdot \frac{b^{(2)}}{\left(b^{(1)}\right)^2}, \tag{71}$$

where

$$E\left[\tilde{W}^{(k)}\right] = \hat{\Lambda}^{-1} \sum_{i=1}^{N} \hat{\lambda}_i E\left[\tilde{W}_i^{(k)}\right], \quad E\left[\left(\tilde{W}^{(k)}\right)^2\right] = \hat{\Lambda}^{-1} \sum_{i=1}^{N} \hat{\lambda}_i E\left[\left(\tilde{W}_i^{(k)}\right)^2\right], \tag{72}$$

and where

$$\sum_{i=1}^{N} \hat{\rho}_i E\left[\tilde{W}_i^{(k)}\right] = \frac{b^{(2)}}{2b^{(1)}} + r\delta_k, \tag{73}$$

which follows directly from the pseudo-conservation law derived in [39]. The following result follows directly from Theorem 1.

Property 4 (Second moment of the delay)

For $i = 1, \ldots, N$, we have for the one-phase gated model

$$E\left[\left(\tilde{W}_i^{(1)}\right)^2\right] = \frac{1 + \hat{\rho}_i + \hat{\rho}_i^2}{3} \left(r + \frac{1}{2\delta_1} \frac{b^{(2)}}{b^{(1)}} \right) \left(r + \frac{1}{\delta_1} \frac{b^{(2)}}{b^{(1)}} \right), \tag{74}$$

and for the two-stage gated model

$$E\left[\left(\tilde{W}_i^{(2)}\right)^2\right] = \frac{7 + 4\hat{\rho}_i + \hat{\rho}_i^2}{3} \left(r + \frac{1}{2\delta_2} \frac{b^{(2)}}{b^{(1)}} \right) \left(r + \frac{1}{\delta_2} \frac{b^{(2)}}{b^{(1)}} \right). \tag{75}$$

An exact expression for the asymptotic SQF (70) of both models can then be obtained by combining (72)-(75). Although an in-depth study of the asymptotic SQF properties of the one-phase and

two-phase models is beyond the scope of the paper, we briefly touch upon some considerations about SQF below.

An interesting question is whether in general the two-phase gated polling model, which is known to be (asymptotically) less efficient, is asymptotically less unfair than its one-phase counterpart (Property 2), which is one of the reasons for proposing the two-phase policy in the first place. The answer is no. To illustrate this, consider the three-queue model with the following parameters: $\hat{\lambda}_1 = 1/10$, $\hat{\lambda}_2 = 2/10$, $\hat{\lambda}_3 = 3/10$, all the service times are deterministic with means 1, 2 and 3, respectively, so that $b^{(1)} = 5/3$ and $b^{(2)} = 10/3$ and $\hat{\rho}_1 = 3/10$, $\hat{\rho}_2 = 4/10$ and $\hat{\rho}_3 = 3/10$. Table 1 below shows the asymptotic unfairness $\hat{\mathcal{F}}^{(k)}_{SQF}$ for $k = 1, 2$ for different values of the mean switch-over time per cycle r. The results in Table 1 illustrate the fact that in general there is no

r	$\hat{\mathcal{F}}^{(1)}_{SQF}$	$\hat{\mathcal{F}}^{(2)}_{SQF}$
0	1.35	1.22
0.05	1.40	1.34
0.1	1.46	1.46
0.25	1.63	1.85
1.0	2.55	4.18

Table 1: Asymptotic customer unfairness for different values of r.

dominance relation between the one-phase and two-phase gated polling systems with respect to the asymptotic unfairness measure defined in (70).

The SQF-measure defined in (70) includes the combined impact of *seniority* and *service-time variability*. Therefore, one might suspect that if all service times are deterministic with the same mean the two-phase gated service is more fair in the SQF-sense, since the effect of overtaking (i.e. customers that arrive earlier are served later) seems to be less. However, in general such a relation does not hold. To illustrate this, consider the following three-queue model: $\hat{\lambda}_1 = 1/12$, $\hat{\lambda}_2 = 1/6$, $\hat{\lambda}_3 = 1/4$, all the service times are deterministic with mean 2, so that $b^{(1)} = 2$ and $b^{(2)} = 4$ and $\hat{\rho}_1 = 1/6$, $\hat{\rho}_2 = 1/3$ and $\hat{\rho}_3 = 1/2$. Table 2 shows the asymptotic unfairness $\hat{\mathcal{F}}^{(k)}_{SQF}$ for $k = 1, 2$ for different values of r. Table 2 illustrates that even if the service-time distributions are deterministic

r	$\hat{\mathcal{F}}^{(1)}_{SQF}$	$\hat{\mathcal{F}}^{(2)}_{SQF}$
0	1.15	1.03
0.1	1.24	1.20
0.2	1.32	1.38
0.3	1.59	1.91
1.0	2.04	2.82

Table 2: Asymptotic customer unfairness for different values of r.

and identical for all queues there is no dominance relation between the one-phase and two-phase gated polling systems with respect to (70).

The results presented in (72)-(75) lead to a closed-form expression for the asymptotic SQF-measure for one-phase and two-phase polling models, and can be extended to a much broader class of MTBP-type polling models, following the general framework developed in [37]. The preliminary results presented in Tables 1 and 2 show that the derivation of dominance relations between fairness of policies with respect to the SQF-measure is far from trivial, and beyond the scope of the present paper, addressing a challenging area for further research.

5 Further Research

The results presented in this paper suggest a number of interesting topics for further research. First, it is a tremendous challenge to develop some type of ordering of service policies with respect to (asymptotic) efficiency and fairness. In the context of efficiency, such an ordering has been obtained in [40]. However, the development of an ordering with respect to the fairness measure defined in (70) [6], where *seniority* and *service time variability* are intertwined, is an open and intriguing problem. The results shown in Table 1 suggest that developing such an ordering is far from trivial. Second, the two-phase gated service policy may be naturally extended to a general K_i-phase gated service policy, where queue i receives K_i-phase gated service, for $i = 1, \ldots, N$. We suspect that the results in the present paper can be extended to the case of K_i-phase gated service by considering the evolution of the system as a $K := \sum_{j=1}^{N} K_j$-dimensional MTBP. Such results would also raise challenging questions regarding the optimal setting of the K_i-values that properly balance fairness and efficiency. Extension of the results presented in this paper form an interesting topic for further research. Third, in this paper it is assumed that the service-time distributions have finite variance. It would be interesting to investigate if the results can be extended to include infinite-variance (e.g., regularly varying) service-time distributions. In this context, note however that Quine's result (Theorem 2) explicitly relies on the finite-variance assumptions and no longer holds if the assumption is violated. Extension of the results to service times with infinite variance would be a breakthrough in the field. Finally, it would be interesting both from a theoretical and application point-of-view to extend the results to non-Poisson arrivals. In this context, we can build upon the recent results presented in [40], where we rigorously prove HT limits for polling models with gated and exhauistive service at all queues and with renewal arrivals .

Acknowledgment: The authors would like to thank the referees for their useful suggestions, which have led to a significant improvement of the paper.

References

[1] Altman, E. and Kushner, H. (2002). Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. *SIAM J. Control and Optimization* **41**, 217-252.

[2] Altman, E. and Fiems, D. (2007). Expected waiting time in symmetric polling systems with correlated vacations. *Queueing Systems* **56**, 241-253.

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

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

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

[6] Brosh, E., Levy, H. and Avi-Itzhak, B. (2007) SQF: A Slowdown Queueing Fairness Measure. *Performance Evaluation* **64**, 1121-1136.

[7] 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.

[8] 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.

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

[10] Groenevelt, R. and Altman, E. (2005). Analysis of alternating-priority queueing models with (cross) correlated switchover times. *Queueing Systems* **51**, 199-247.

[11] 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.

[12] Kramer, G., Mukherjee, B. and Pesavento, G. (2001). Ethernet PON: design and analysis of an optical access network. *Phot. Net. Commun.* **3**, 307-319.

[13] Kramer, G., Mukherjee, B. and Pesavento, G. (2002). Interleaved polling with adaptive cycle time (IPACT): a dynamic bandwidth allocation scheme in an optical access network. *Phot. Net. Commun.* **4**, 89-107.

[14] Kramer, G., Mukherjee, B. and Pesavento, G. (2002). Supporting differentiated classes of services in Ethernet passive optical networks. *J. Opt. Netw.* **1**, 280-290.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Olsen, T.L. and Van der Mei, R.D. (2005). Periodic polling systems in heavy-traffic: renewal arrivals. *Operations Research Letters* **33**, 17-25.

[23] Olsen, T.L.(2001). Limit theorems for polling models with increasing setups. *Prob. Eng. Inf. Syst.* **15**, 35-55.

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

[25] Raz, D., Levy, H. and Avi-Ithzak, B. (2003). A resource allocation queueing fairness measure. In: *Proc. ACM Sigmetrics* (San Diego, CA), 130-141.

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

[27] 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.

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

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

[30] 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.

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

[32] 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.

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

[34] Van der Mei, R.D. (1999). Delay in polling systems with large switch-over times. *Journal of Applied Probability* **36**, 232-243.

[35] Van der Mei, R.D. (1999). Polling systems in heavy traffic: higher moments of the delay. *Queueing Systems* **31**, 365-394.

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

[37] Van der Mei, R.D. (2008). Towards a unifying theory on branching-type polling models in heavy traffic. To appear in *Queueing Systems*.

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

[39] Van der Mei, R.D. and Resing, J.A.C. (2007). Analysis of polling systems with two-stage gated service: fairness versus efficiency. In: Managing Traffic Performance in Converged Networks - the Interplay between Convergent and Divergent Forces (eds. L. Mason, T. Drwiega and J. Yan), ITC2007, Lecture Notes in Computer Science 4516, 544-555.

[40] Van der Mei, R.D. and Winands, E.M.M. (2007). Polling models with renewal arrivals: a new method to derive heavy-traffic asymptotics. *Performance Evaluation* **64**, 1029-1040.

[41] Van der Mei, R.D. and Winands, E.M.M. (2008). Mean Value Analysis for polling models in heavy traffic. To appear in *Performance Evaluation*.

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

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

[44] Wierman, A. and Harchol-Balter, M. (2003). Classifying scheduling policies with respect to unfairness in an M/G/1. In: *Proc. ACM Sigmetrics* (San Diego, CA), 238-249.

[45] Winands, E.M.M., Adan, I.J.B.F. Adan and Van Houtum, G.J. (2006). Mean value analysis for polling systems. *Queueing Systems* **54**, 35-44.

[46] Winands, E.M.M. (2006). Branching-type polling systems with large setups. Technical Report, Technische Universiteit Eindhoven.

Appendix A: The Descendant Set Approach for the Two-Phase Gated Model

In this Appendix we discuss how the model defined in Section 2 can be analyzed by means of the Descendant Set Approach (DSA), introduced in [11] for models with exhaustive and gated service.

The customers in a polling system can be classified as *originators* and *non-originators*. An originator is a customer that arrives at the system during a switch-over period. A non-originator is a customer that arrives at the system during the service of another customer. For a customer C, define the children set to be the set of customers arriving during the service of C; the descendant set of C is recursively defined to consist of C, its children and the descendants of its children. The DSA is focused on the determination of the moments of the delay at a fixed queue, say Q_1. To this end, the DSA concentrates on the determination of the distribution of the two-dimensional stochastic vector $\left(X_1^{(1)}(P^*), X_1^{(2)}(P^*)\right)$, where $X_1^{(k)}(P^*)$ is defined as the number of phase-k customers at Q_1 present at an arbitrary fixed polling instant P^* at Q_1 ($k = 1, 2$). P^* is referred to as the *reference point*. The main ideas are the observations that (1) each of the customers present at Q_1 at the reference point P^* (either at phase 1 or phase 2) belongs to the descendant set of exactly one originator, and (2) the evolutions of the descendant sets of different originators are stochastically independent. Therefore, the DSA concentrates on an arbitrary tagged customer which arrived at Q_i in the past and on calculating the number of type-1 descendants it has at both phases at P^*. Summing up these numbers over all past originators yields $\left(X_1^{(1)}(P^*), X_1^{(2)}(P^*)\right)$, and hence $\left(X_1^{(1)}, X_1^{(2)}\right)$, because P^* is chosen arbitrarily.

The DSA considers the Markov process embedded at the polling instants of the system. To this end, we number the successive polling instants as follows. Let $P_{N,0}$ be the last polling instant at Q_N prior to P^*, and for $i = N-1, \ldots, 1$, let $P_{i,0}$ be recursively defined as the last polling instant at Q_i prior to $P_{i+1,0}$. In addition, for $c = 1, 2, \ldots$, we define $P_{i,c}$ to be the last polling instant at Q_i prior to $P_{i,c-1}$, $i = 1, \ldots, N$. Define the c-th cycle to be the time between $P_{1,c}$ and $P_{1,c-1}$, for $c = 0, 1, \ldots$. The DSA is oriented towards the determination of the contribution to $\left(X_1^{(1)}(P^*), X_1^{(2)}(P^*)\right)$ of an arbitrary customer present at Q_i at $P_{i,c}$. To this end, define an (i,c)-customer to be a customer present at Q_i at $P_{i,c}$. Moreover, for a tagged (i,c)-customer $T_{i,c}$ *at phase-1*, we define $\underline{A}_{i,c} := \left(A_{i,c}^{(1)}, A_{i,c}^{(2)}\right)$, where $A_{i,c}^{(k)}$ is the number of type-1 descendants it has at phase-k at P^* ($k = 1, 2$). In this way, the two-dimensional random variable $\underline{A}_{i,c}$ can be viewed as the contribution of $T_{i,c}$ to $\left(X_1^{(1)}(P^*), X_1^{(2)}(P^*)\right)$. Denote the joint PGF of $\underline{A}_{i,c}$ by, for $|z_1|, |z_2| \leq 1$, $i = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$A_{i,c}^*(z_1, z_2) := E\left[z_1^{A_{i,c}^{(1)}} z_2^{A_{i,c}^{(2)}}\right]. \tag{76}$$

To express the distribution of $\left(X_1^{(1)}, X_1^{(2)}\right)$ in terms of the distributions of the descendant set variables $\underline{A}_{i,c}$, denote by $R_{i,c}$ the switch-over period from Q_i to Q_{i+1} immediately after the service period at Q_i starting at $P_{i,c}$. Moreover, denote $\underline{S}_{i,c} := \left(S_{i,c}^{(1)}, S_{i,c}^{(2)}\right)$, where $S_{i,c}^{(k)}$ is the total contribution to $X_1^{(k)}$ of all customers that arrive at the system during $R_{i,c}$ (note that, by definition, these customers are original customers), and denote the joint PGF of $\underline{S}_{i,c}$ by, $|z_1|, |z_2| \leq 1$, $i = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$S_{i,c}^*(z_1, z_2) := E\left[z_1^{S_{i,c}^{(1)}} z_2^{S_{i,c}^{(2)}}\right]. \tag{77}$$

In this way, $\underline{S}_{i,c} = \left(S_{i,c}^{(1)}, S_{i,c}^{(2)}\right)$ can be seen as the (joint) contribution of $R_{i,c}$ to $\left(X_1^{(1)}(P^*), X_1^{(2)}(P^*)\right)$. It is readily verified that we can write

$$\underline{X}_1 = \left(X_1^{(1)}, X_1^{(2)}\right) = \sum_{i=1}^{N}\sum_{c=0}^{\infty}\left(S_{i,c}^{(1)}, S_{i,c}^{(2)}\right) = \sum_{i=1}^{N}\sum_{c=0}^{\infty}\underline{S}_{i,c}. \tag{78}$$

Note that $S_{i,c}^{(1)}$ and $S_{i',c'}^{(2)}$ are dependent if $(i,c) = (i',c')$ but independent otherwise. Hence we can write, for $|z_1|, |z_2| \leq 1$,

$$X_1^*(z_1,z_2) = \prod_{i=1}^{N}\prod_{c=0}^{\infty} S_{i,c}^*(z_1,z_2). \tag{79}$$

Because $\underline{S}_{i,c}$ is the total joint contribution to $\underline{X}_1$ of all (original) customers that arrive during $R_{i,c}$, the joint distribution of $\underline{S}_{i,c}$ can be expressed in terms of the distributions of the DS-variables $\underline{A}_{i,c}$ as follows: For $i = 1,\ldots,N$, $c = 0,1,\ldots$, and $|z_1|, |z_2| \leq 1$,

$$S_{i,c}^*(z_1,z_2) = R_i^*\left(\sum_{j=i+1}^{N}\left[\lambda_j - \lambda_j A_{j,c}^*(z_1,z_2)\right] + \sum_{j=1}^{i}\left[\lambda_j - \lambda_j A_{j,c-1}^*(z_1,z_2)\right]\right). \tag{80}$$

To define a recursion for the evolution of the descendant set, note that a customer at phase-1 present at Q_1 at the polling instant at Q_1 during cycle c is served during the *next* cycle, which leads to the following relation: For $i = 1,\ldots,N$, $c = 0,1,\ldots$, and $|z_1|, |z_2| \leq 1$,

$$A_{i,c}^*(z_1,z_2) = B_i^*\left(\sum_{j=i+1}^{N}\left[\lambda_j - \lambda_j A_{j,c-1}^*(z_1,z_2)\right] + \sum_{j=1}^{i}\left[\lambda_j - \lambda_j A_{j,c-2}^*(z_1,z_2)\right]\right), \tag{81}$$

supplemented with the basis for the recursion

$$A_{i,-1}^*(z_1,z_2) = z_1 I_{\{i=1\}}, \quad \text{and} \quad A_{i,-2}^*(z_1,z_2) = z_2 I_{\{i=1\}}. \tag{82}$$

In this way, relations (78)-(82) give a complete characterization of the simultaneous distribution of $\left(X_1^{(1)}, X_1^{(2)}\right)$. Similarly, recursive relations to calculate the (cross-)moments of $\left(X_1^{(1)}, X_1^{(2)}\right)$ can be readily obtained from those equations.

Relation (81) leads to the following recursive relations for the first moment of the DS variables $A_{i,c}^{(k)}$. More precisely, if we define for $i = 1,\ldots,N$, $c = -2,-1,0,1,\ldots$ and $k = 1,2$,

$$\alpha_{i,c}^{(k)} := E\left[A_{i,c}^{(k)}\right], \tag{83}$$

then (78)-(82) are easily seen to lead to the following recursive scheme: For $i = 1,\ldots,N$, $c = 0,1,\ldots$, and $k = 1,2$,

$$\alpha_{i,c}^{(k)} = b_i^{(1)}\left[\sum_{j=i+1}^{N}\lambda_j\alpha_{j,c-1}^{(k)} + \sum_{j=1}^{i}\lambda_j\alpha_{j,c-2}^{(k)}\right], \tag{84}$$

supplemented with the following basis for the recursion, for $i = 1,\ldots,N$,

$$\alpha_{i,-2}^{(1)} := 0, \ \alpha_{i,-2}^{(2)} := I_{\{i=1\}}, \ \alpha_{i,-1}^{(1)} := I_{\{i=1\}} \text{ and } \alpha_{i,-1}^{(2)} := 0. \tag{85}$$

Note that since a phase-1 customer at Q_i present in the system at $P_{i,c}$ is served during the $(c-1)$-st cycle, the contribution of that phase-1 customer is stochastically identical to that of a

type-i customer in phase-2 who is present at the system at $P_{i,c-1}$. Consequently, we have, for $i = 1, \ldots, N$, $c = -1, 0, 1, \ldots$, that

$$\alpha_{i,c}^{(1)} = \alpha_{i,c-1}^{(2)}. \tag{86}$$

Moreover, it follows directly from (78) and (80) that, for $k = 1, 2$,

$$E\left[X_1^{(k)}\right] = \sum_{i=1}^{N}\sum_{c=0}^{\infty} E\left[S_{i,c}^{(k)}\right] = \sum_{i=1}^{N}\sum_{c=0}^{\infty} r_i^{(1)}\left[\sum_{j=i+1}^{N}\lambda_j\alpha_{j,c}^{(k)} + \sum_{j=1}^{i}\lambda_j\alpha_{j,c-1}^{(k)}\right]. \tag{87}$$