Polling Models with Multi-Phase Gated Service

R.D. van der Mei[a,b] and A. Roubos[b]

[a]Centre for Mathematics and Computer Science, Amsterdam, The Netherlands
[b]Department of Mathematics, VU University Amsterdam, The Netherlands

January 16, 2011

Abstract

In this paper we introduce and analyze a new class of service policies called multi-phase gated service. This policy is a generalization of the classical single-phase and two-phase gated policies and works as follows. Each customer that arrives at queue i will have to wait $K_i \geq 1$ cycles before it receives service. The aim of this policy is to provide an interleaving scheme to avoid monopolization of the system by heavily loaded queues, by choosing the proper values of interleaving levels K_i. In this paper, we analyze the effectiveness of the interleaving scheme on the queueing behavior of the system, and consider the problem of identifying the proper combination of interleaving levels $\underline{K}^* = (K_1^*, \ldots, K_N^*)$ that minimizes a weighted sum of the mean waiting times at each of the N queues. Obviously, the proper choice of the interleaving levels is most critical when the system is heavily loaded. For this reason, we explore the framework developed in [26] to obtain closed-form expressions for the asymptotic waiting-time distributions in heavy traffic, and use these expressions to derive simple heuristics for approximating the optimal interleaving scheme $\underline{K}^*$. Numerical results with simulations demonstrate that the accuracy of these approximations is extremely high.

1 Introduction

This study is motivated by dynamic bandwidth allocation schemes 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 voice, 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 [14]. 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. [15, 16] 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.

In the present paper, we analyze and optimize the efficiency of interleaving schemes by modeling the bandwidth sharing between the ONUs by cyclic polling models with a multi-phase gated service policy. In this policy, a customer that arrives at queue i has to wait K_i cycles before it can be taken into service. This interleaving scheme, which is a natural extension of the classical (one-phase) gated and the two-phase gated service policy studied in [28, 20], provides the flexibility to properly manage the relative waiting times among the queues by tuning the interleaving levels $\underline{K} = (K_1, \ldots, K_N)$, opening up possibilities for performance improvement and optimization.

Exact analysis of the delay in polling systems 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 systems 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, which leads to significant insights in the behavior of the system (e.g., insensitivity and monotonicity properties). Secondly, the efficiency of the numerical algorithms tends to degrade 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 attractiveness of using heavy-traffic asymptotics as the basis for optimization of the system performance.

We consider an asymmetric Poisson-driven cyclic polling model with N queues and with generally distributed service times and switch-over times. Each queue receives multi-phase gated service with parameters $\underline{K} = (K_1, \ldots, K_N)$, which works as follows. Newly incoming customers are first queued at the phase-1 buffer. When the server arrives at queue i, it closes the gate behind the customers residing in the phase-1 buffer, then serves all customers waiting in the phase-K_i buffer on a First-Come-First-Served (FCFS) basis, and moves all customers before the gate at the phase-j buffer to the phase-$(j+1)$ buffer, for $j = 1, \ldots, K_i - 1$, before moving to the next queue. The K_i-phase gated service policy was introduced by Park et al. [20] for the case $K_i = 2$ $(i = 1, \ldots, N)$. The model under consideration is easily seen to have a Multitype Branching Process (MTBP) structure [22]. Exploring this structure, Van der Mei and Resing [28] derived closed-form expressions for the complete asymptotic distributions of the waiting-times in the two-phase gated polling model, when the load ρ tends to unity, under proper heavy-traffic scalings. Recently, Van der Mei [26] developed a framework for deriving heavy-traffic asymptotics for a general class of MTBP-type of polling models. In this paper, we apply this framework to multi-phase polling models to obtain a closed-form expression for the Laplace-Stieltjes Transform (LST) of the limiting distribution of $(1-\rho)W_i$ $(i = 1, \ldots, N)$ as ρ goes to 1, where W_i is the waiting time at queue i. We also give strong conjectures for heavy-traffic asymptotics for renewal arrivals. The expressions are strikingly simple and show explicitly how the waiting-time distributions depend on the system parameters, and in particular, on the interleaving levels K_i $(i = 1, \ldots, N)$. These asymptotic results directly lead to a number of asymptotic insensitivity properties of the waiting-time distributions with respect to the system parameters, and moreover, lead to simple approximations for the moments and tail probabilities of the waiting times. Numerical results are presented to assess the accuracy of these approximations. Finally, we consider the problem of finding a combination $\underline{K}^* = (K_1^*, \ldots, K_N^*)$ of interleaving levels that minimizes $\sum_{i=1}^{N} c_i E[W_i]$, where the weights $c_1, \ldots, c_N \geq 0$ can be chosen arbitrarily. Using the asymptotic results, we propose simple heuristics. Numerical results show that the heuristics lead to excellent results.

The results in this paper generalize those in [28], where the special case $K_i = 2$ $(i = 1, \ldots, N)$ was considered. The contribution of the present paper compared to [28] is three-fold. First, we propose a new class of service policies that are motivated by DBA problems for EPONs. This class is attractive because it is flexible and allows for optimization of the system performance. Second, we derive new and simple closed-form expressions for the asymptotic distribution and moments of the scaled waiting-times in heavy traffic, which provides new and valuable insight in how the system performs as a function of $\underline{K}$. Third, we use these asymptotics to propose and validate simple heuristics for the "optimal" choice of $\underline{K}$.

The remainder of this paper is organized as follows. In Section 2 the model is described. In Section 3 we derive a pseudo-conservation law for the model, and present asymptotic expressions for the waiting-time distribution in heavy traffic, and a number of asymptotic insensitivity properties. These expressions suggest simple approximations for the moments and tail probabilities of the waiting times for stable systems; in Section 4 these approximations are validated. In Section 5 we focus on optimization. To this end, we use the asymptotic results to develop simple heuristics for determining the combination of interleaving levels $\underline{K}^*$ that minimizes a weighted sum of the mean waiting times. The heuristics are validated extensively by simulations. In Appendix A we give a general description of MTBPs, used for reference. In Appendix B we discuss the details of the use of the Descendant Set Approach (DSA), a numerical technique to calculate the moments and tail probabilities of the waiting times at each of the queues. Finally, in Appendix C we use the stepwise approach in [28] to prove the heavy-traffic results.

2 Model Description

Consider a system consisting of $N \geq 2$ queues $Q_1, \ldots, Q_N$. Q_i consists of $K_i \geq 1$ buffers: a phase-1 buffer, a phase-2 buffer up to a phase-K_i buffer, $i = 1, \ldots, N$. Let $K := \sum_{i=1}^{N} K_i$. 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 LST $B_i^*(s)$ and with finite k-th moment $b_i^{(k)}$ $(k = 1, 2, \ldots)$. The k-th moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \sum_{i=1}^{N} \lambda_i b_i^{(k)} / \Lambda$ $(k = 1, 2, \ldots)$. The load offered to Q_i is $\rho_i = \lambda_i b_i^{(1)}$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$. 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 K_i-phase gated service policy, which works as follows. When the server arrives at Q_i, it closes the gate behind the customers residing in the phase-1 buffer. Then, all customers waiting in the phase-K_i buffer are served on a FCFS basis. Subsequently, all customers before the gate at the phase-k buffer are instantaneously forwarded to the phase-$(k+1)$ buffer $(k = 1, \ldots, K_i - 1)$, and the server proceeds to the next queue. Denote $\underline{K} := (K_1, \ldots, K_N)$, and denote the set of possible values of $\underline{K}$ by $\mathcal{S} := \{1, 2, \ldots\}^N$. Upon departure from Q_i the server immediately proceeds to Q_{i+1}, incurring a switch-over time R_i, with LST $R_i^*(s)$ and finite k-th moment $r_i^{(k)}$ $(k = 1, 2, \ldots)$. Moreover, 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, and denote the second moment by $r^{(2)} = \sum_{i=1}^{N} r_i^{(2)} + \sum_{i \neq j} r_i^{(1)} r_j^{(1)}$. 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. [11]). Let W_i be 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, and denote thee corresponding LST by $W_i^*(s)$.

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

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

and Laplace-Stieltjes Transform (LST)

$$\Gamma^*(s) = \left(\frac{\mu}{\mu + s} \right)^\alpha \quad (Re(s) > 0). \tag{2}$$

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

The heavy-traffic limits, denoted $\rho \uparrow 1$, taken in this paper are such that the arrival rates are increased, while keeping both the service-time distributions and the ratios between the arrival rates fixed. For a random vector $\underline{Y}$ we denote $(1-\rho)\underline{Y} \to_d \underline{\tilde{Y}}$ $(\rho \uparrow 1)$ if for all $\epsilon > 0$ there exist $\delta > 0$ such that if $|1-\rho| < \delta$ then it holds that

$$\sup_{\underline{y}} \left| \Pr\left\{(1-\rho)\underline{Y} < \underline{y}\right\} - \Pr\left\{\underline{\tilde{Y}} < \underline{y}\right\} \right| < \epsilon. \tag{3}$$

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. Denote by $\underline{1}$ a vector whose entries are all 1. 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. A K-dimensional vector $\underline{x}$ has components $\underline{x} = (x_1^{(1)}, \ldots, x_1^{(K_1)}, \ldots, x_N^{(1)}, \ldots, x_N^{(K_N)})$. Finally, the notation $[\cdot]$ means rounding off to the nearest positive integer.

3 Analysis

In Section 3.1 we present a pseudo-conservation law for the model under study. In Section 3.2 we derive some preliminary results that will be used in Section 3.3 to derive heavy-traffic limits for the waiting-time distributions at each of the queues. We also formulate conjectures for extension of the results to renewal arrivals.

3.1 Pseudo-Conservation Law

On the basis of the principle of work decomposition, we have (cf. [5]): For $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i E[W_i] = \rho \frac{\rho}{1-\rho} \frac{b^{(2)}}{2b^{(1)}} + \rho \frac{r^{(2)}}{2r} + \frac{r}{2(1-\rho)} \left[\rho^2 - \sum_{i=1}^{N} \rho_i^2 \right] + \sum_{i=1}^{N} E\left[M_i\right], \tag{4}$$

where M_i stands for the amount of work at Q_i at an arbitrary moment at which the server departs from Q_i. It is clear that $M_i = M_i^{(1)} + \cdots + M_i^{(K_i)}$, where $M_i^{(k)}$ is the amount of work at phase k at a server departure epoch from Q_i $(k = 1, \ldots, K_i)$. Then simple balancing arguments can be used to show that $E[M_i^{(1)}] = \rho_i^2 r/(1-\rho)$, and for $k = 2, \ldots, K_i$, $E[M_i^{(k)}] = \rho_i r/(1-\rho)$, which immediately implies

$$E\left[M_i\right] = \rho_i \left((K_i - 1) + \rho_i\right) \frac{r}{1-\rho}. \tag{5}$$

3.2 Preliminaries

In this section we derive expressions for the asymptotic waiting-time distributions when the load ρ approaches 1. To this end, the following notation is convenient. Without loss of generality, we 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 present at Q_i *at an arbitrary polling instant at* Q_1 when the system is in steady state $(k = 1, \ldots, K_i,\ i = 1, \ldots, N)$. Moreover, for $i = 1, \ldots, N$, define the K_i-dimensional random variable

$$\underline{X}_i := (X_i^{(1)}, \ldots, X_i^{(K_i)}), \text{ with joint PGF } \tilde{X}_i^*(z_1, \ldots, z_{K_i}). \tag{6}$$

Denote by $X_1 := X_1^{(1)} + \cdots + X_1^{(K_1)}$ the *total* number of customers at Q_1 *at the beginning* of a visit period to Q_1, and denote the corresponding (one-dimensional) PGF by $X_1^*(z)$. Similarly, denote

by Y_1 the total number of customers at Q_1 *at the end* of a visit period to Q_1, and denote the corresponding PGF by $Y_1^*(z)$. Then it is easily verified that, for $|z| \leq 1$,

$$X_1^*(z) = \tilde{X}_1^*(z, \ldots, z, z), \quad \text{and} \quad Y_1^*(z) = \tilde{X}_1^*(z, \ldots, z, B_1^*(\lambda_1(1-z))). \tag{7}$$

Also denote by N_1 the total number of customers at Q_1 (possibly including a customer in service), *at an arbitrary moment*, and denote the corresponding PGF by $N_1^*(z)$. Recall that W_i is the waiting time of an arbitrary customer at Q_i, with corresponding LST $W_i^*(s)$. Moreover, let S_i be the sojourn time of an arbitrary type-i customer in the system, and denote the corresponding LST by $S_i^*(s)$. Note that since W_i and B_i are independent, it holds that, for $Re(s) \geq 0$,

$$S_i^*(s) = W_i^*(s) B_i^*(s). \tag{8}$$

Moreover, by applying the distributional for of Little's Law it is readily seen that, for $Re(s) \geq 0$,

$$S_1^*(s) = N_1^*(1 - s/\lambda_1). \tag{9}$$

Then the following result gives an expression for the LST of W_1 in terms of the distribution of the K_1-dimensional random variable $\underline{X}_1$, defined in (6).

Lemma 1
For $\rho < 1$, $Re(s) \geq 0$,

$$W_1^*(s) = \frac{\tilde{X}_1^*(1 - s/\lambda_1, \ldots, 1 - s/\lambda_1, B_1^*(s)) - \tilde{X}_1^*(1 - s/\lambda_1, \ldots, 1 - s/\lambda_1, 1 - s/\lambda_1)}{E\left[X_1^{(1)}\right](B_1^*(s) - 1 + s/\lambda_1)}. \tag{10}$$

Proof: The following result gives a relation between N_1, X_1 and Y_1 (cf. [4]): For $|z| \leq 1$,

$$N_1^*(z) = \frac{B_1^*(\lambda_1(1-z))}{E\left[X_1^{(1)}\right]} \cdot \frac{Y_1^*(z) - X_1^*(z)}{B_1^*(\lambda_1(1-z)) - z}. \tag{11}$$

The result follows then directly by taking $z := 1 - s/\lambda_1$, and using (7), (8) and (9). $\square$

Straightforward balancing arguments lead to the following expression for the first moment $E[X_1^{(k)}]$: For $\rho < 1$, $k = 1, \ldots, K_1$,

$$E[X_1^{(k)}] = \frac{\lambda_1 r}{1 - \rho}. \tag{12}$$

In general, the higher-order moments and the distributions of $X_1^{(k)}$ cannot be obtained explicitly for arbitrary values of the load, but can be calculated by numerical techniques, such as the classical buffer-occupancy approach [23], and the Descendant Set Approach (DSA) [13, 8]. We refer to [28] for a discussion of the use of the DSA for the special case $K_i = 2$ ($i = 1, \ldots, N$), and to [31] for an overview of the solution techniques for polling models. Throughout, the following notation is useful. Let

$$\underline{X} := \left(X_1^{(1)}, \ldots, X_1^{(K_1)}, \ldots, X_N^{(1)}, \ldots, X_N^{(K_N)}\right) \tag{13}$$

be the $K = \sum_{j=1}^{N} K_j$-dimensional vector that describes the state of the system at an arbitrary polling instant at Q_1. Recall that $X_i^{(k)}$ stands for the steady-state number of phase-k customers that reside at Q_i *at an arbitrary polling instant at* Q_1. To determine the asymptotic behavior of W_1, Equation (10) implies that it suffices to determine the limiting behavior of $\underline{X}_1$, defined in (6), as ρ tends to 1. To this end, we first show how the evolution of the system at successive polling instants at Q_1 can be described as a MTBP with immigration in each state. Then, in the next section we use this MTBP-description to derive an asymptotic expression for the limiting

distribution of $\underline{X}$, and hence for $\underline{X}_1$ by taking the first K_1 entries in (13) only, as $\rho \uparrow 1$.

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

$$\underline{X}^{(n)} := \left(X_{1,n}^{(1)}, \ldots, X_{1,n}^{(K_1)}, \ldots, X_{N,n}^{(1)}, \ldots, X_{N,n}^{(K_N)} \right) \tag{14}$$

be the state vector at the n-th polling instant at Q_1. Then similar to the analysis made by Resing [22] we make the following observation.

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

$$f(\underline{s}) := \left(f^{(1,1)}(\underline{s}), \ldots, f^{(1,K_1)}(\underline{s}), \ldots, f^{(N,1)}(\underline{s}), \ldots, f^{(N,K_N)}(\underline{s}) \right), \tag{15}$$

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

$$f^{(i,k)}(\underline{s}) := s_i^{(k+1)} \text{ for } k = 1, \ldots, K_i - 1, \tag{16}$$

and

$$f^{(i,K_i)}(\underline{s}) := B_i^* \left(\sum_{j=1}^{i} \lambda_j \left(1 - s_j^{(1)} \right) + \sum_{j=i+1}^{N} \lambda_j \left(1 - f^{(j,1)}(\underline{s}) \right) \right), \tag{17}$$

and where the PGF of the immigration function is given by

$$g(\underline{s}) := \prod_{i=1}^{N} R_i^* \left(\sum_{j=1}^{i} \lambda_j \left(1 - s_j^{(1)} \right) + \sum_{j=i+1}^{N} \lambda_j \left(1 - f^{(j,1)}(\underline{s}) \right) \right). \tag{18}$$

Proof: Relations (15)–(18) can be obtained along the lines of [28] for the case of two-phase gated service, using simple generating-function manipulations. More specifically, in the spirit of the work in [28], equation (16) follows from the fact that for $k = 1, \ldots, K_i - 1$ it holds that a type-i customer at phase k at a given polling instant P_1 at Q_1 is not served during the visit period starting at P_1, but is simply forwarded from phase k to phase $k + 1$. In this way, this customer is "effectively replaced" by a single type-i customer at phase $k + 1$ at the next polling instant at Q_1. Similarly, (17) follows from the fact that each type-i customer at phase K_1 at P_1 is served during the visit period starting at P_1, and hence, is "effectively replaced" by all customers that arrive in the system during its service time with LST $B_i^*(\cdot)$. Finally, (18) 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 $R_i^*(\cdot), i = 1, \ldots, N$. $\square$

Remark 1. Theorem 1 corresponds to the results in [22] for the case of single-phase gated service at all queues (i.e., $K_i = 1$ for all i). Note that in that case it holds that $K = N$ and $\underline{s} = (s_1^{(1)}, \ldots, s_N^{(1)})$, and that (16) disappears, while (17) simply becomes: For $i = 1, \ldots, N$,

$$f^{(i)}(\underline{s}) = B_i^* \left(\sum_{j=1}^{i} \lambda_j (1 - s_j^{(1)}) + \sum_{j=i+1}^{N} \lambda_j (1 - f^{(1,1)}(\underline{s})) \right), \tag{19}$$

and Equation (18) simplifies to

$$g(\underline{s}) = \prod_{i=1}^{N} R_i^* \left(\sum_{j=1}^{i} \lambda_j (1 - s_j^{(1)}) + \sum_{j=i+1}^{N} \lambda_j (1 - f^{(j,1)}(\underline{s})) \right). \tag{20}$$

Similarly, for $K_i = 2$ $(i = 1, \ldots, N)$ the results are in line with the ones in [28].

For later reference, it is convenient to define the mean offspring matrix as follows: For $i,j = 1,\ldots,N,\ k=1,\ldots,K_i,\ l=1,\ldots,K_j$,

$$m_{i,j}^{(k,l)} := \left| \frac{\partial}{\partial s_j^{(l)}} f^{(i,k)}(\underline{s}) \right|_{\underline{s}=\underline{1}}, \tag{21}$$

i.e., the mean number of "children" that a type-i customer at phase k has of type-j at phase l in the MTBP defined in Theorem 1. Using this definition, the mean offspring matrix $\mathbf{M}$ can be expressed as the block matrix:

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{1,1} & \cdots & \mathbf{M}_{1,N} \\ \vdots & \vdots & \vdots \\ \mathbf{M}_{N,1} & \cdots & \mathbf{M}_{N,N} \end{pmatrix}, \text{ with } \mathbf{M}_{i,j} = \begin{pmatrix} m_{i,j}^{(1,1)} & \cdots & m_{i,j}^{(1,K_j)} \\ \vdots & \vdots & \vdots \\ m_{i,j}^{(K_i,1)} & \cdots & m_{i,j}^{(K_i,K_j)} \end{pmatrix}, \tag{22}$$

for $i,j=1,\ldots,N$. Then it follows directly from (16), (17) and (21) that, for $i,j=1,\ldots,N$ and $l=1,\ldots,K_j$,

$$m_{i,j}^{(k,l)} = I_{\{l=k+1,\ i=j\}} \quad (k=1,\ldots,K_i-1) \tag{23}$$

and

$$m_{i,j}^{(K_i,l)} = b_i^{(1)} \left[\lambda_j I_{\{l=1,\ j\le i\}} + \lambda_j m_{j,j}^{(1,l)} I_{\{j>i\}} \right]. \tag{24}$$

These results will be useful throughout to derive the heavy-traffic asymptotics.

3.3 Heavy-traffic asymptotics

The following result characterizes the limiting behavior of the state vector when the load goes to 1.

Theorem 2. *The state vector at polling instants at Q_1 satisfies the following asymptotic behavior:*

$$(1-\rho) \begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_1^{(K_1)} \\ \vdots \\ X_N^{(1)} \\ \vdots \\ X_N^{(K_N)} \end{pmatrix} \to \delta \cdot A \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_K \end{pmatrix} \Gamma(\alpha,1) \quad (\rho \uparrow 1), \tag{25}$$

where

$$\alpha = r\delta \frac{b^{(1)}}{b^{(2)}}, \quad A = |\underline{b}|^{-1} \delta^{-1} \frac{b^{(2)}}{2b^{(1)}} \quad \textit{and} \quad \delta = \frac{1}{2} \sum_{i=1}^{N} \hat{\rho}_i((2K_i-1)+\hat{\rho}_i), \tag{26}$$

and where $\underline{\hat{v}}$ is the left eigenvector of the mean matrix $\mathbf{M}$ at $\rho=1$, characterized in Lemmas C.1 and C.2 in Appendix C.

Proof: The derivation of (25)–(26) is a natural extension of the proof for the special case of two-phase gated service in [28], and can be obtained following the stepwise approach proposed in [26], exploring the MTBP-structure of the model and using the Descendant Set Approach (DSA). The details are outlined in Appendices A, B and C. □

Remark 2. Recall from equation (3) that the convergence in (25) should be interpreted 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

$$\sup_{\underline{x} \in \mathcal{R}^K} \left| \text{Prob}\left\{(1-\rho)\underline{X}^{(n)} \leq \underline{x}\right\} - \text{Prob}\left\{\delta \cdot A \cdot \Gamma(\alpha, 1) \cdot \underline{\hat{v}} \leq \underline{x}\right\} \right| < \epsilon, \tag{27}$$

where $\underline{X}^{(n)}$ is defined in (14).

Theorem 3. *Let $\tilde{W}_i$ be a non-negative random variable with LST*

$$\tilde{W}_i^*(s) = \frac{1}{(1-\hat{\rho}_i)rs} \left\{ \left(\frac{\mu}{\mu + (K_i - 1 + \hat{\rho}_i)s} \right)^{\alpha} - \left(\frac{\mu}{\mu + K_i s} \right)^{\alpha} \right\} \quad (Re(s) > 0), \tag{28}$$

with

$$\alpha = 2r\delta \frac{b^{(1)}}{b^{(2)}}, \quad \mu = 2\delta \frac{b^{(1)}}{b^{(2)}} \quad \textit{and} \quad \delta = \frac{1}{2} \sum_{j=1}^{N} \hat{\rho}_i((2K_i - 1) + \hat{\rho}_i). \tag{29}$$

Then for the multi-phase gated model, the distribution of W_i satisfies the following limiting behavior: For $i = 1, \ldots, N$,

$$(1-\rho)W_i \rightarrow_d \tilde{W}_i \quad (\rho \uparrow 1). \tag{30}$$

Proof. The results can be obtained by combining Theorem 2 and Lemma 1, following the lines similar to those in [28] for the case $K_i = 2$ $(i = 1, \ldots, N)$. The details are omitted for compactness. Note that the specifics of the convergence have been defined in (27). □

Theorem 3 reveals a variety of properties about the dependence of the asymptotic delay distribution with respect to the system parameters.

Corollary 1 (Insensitivity)**.** *For $i = 1, \ldots, N$, the distribution of $\tilde{W}_i$*

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 service-time distributions only through $b^{(1)}$ and $b^{(2)}$, i.e., the first and second moment of the service time of an arbitrary customer.*

Corollary 1 is known to be not generally valid for waiting-time distributions in 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, Corollary 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.

Corollary 2 (Zero switch-over times)**.** *For the case of zero switch-over times, the LST of $\tilde{W}_i$ is given by the following expression: For $i = 1, \ldots, N$, $Re(s) \geq 0$,*

$$\lim_{r \downarrow 0} \tilde{W}_i^*(s) = \frac{\delta}{(1-\hat{\rho}_i)s} \frac{b^{(1)}}{b^{(2)}} \log\left(\frac{\mu + K_i s}{\mu + s(K_i - 1 + \hat{\rho}_i)} \right), \tag{31}$$

where δ and μ are defined in (29)*, and where* $\log(\cdot)$ *is an inverse function of the (complex) function* $l(z) := \exp(z)$.

This result follows directly from Theorem 3 by taking the limit for $r \downarrow 0$ in (28), by using (29) and standard algebraic manipulations.

Corollary 3 (Moments of the asymptotic delay)**.** *The k-th moment of the asymptotic delay at Q_i is given by the following expression: For $i = 1, \ldots, N$, $k = 1, 2, \ldots$,*

$$E[\tilde{W}_i^k] = \frac{K_i^{k+1} - (K_i - 1 + \hat{\rho}_i)^{k+1}}{(k+1)(1-\hat{\rho}_i)} \prod_{m=1}^{k} \left[r + m \frac{b^{(2)}/b^{(1)}}{\sum_{j=1}^{N} \hat{\rho}_j (2K_j - 1 + \hat{\rho}_j)} \right]. \tag{32}$$

This result follows directly from Theorem 3 by taking the k-th order derivative in (28), letting $s \to 0$, using (29) and by standard algebraic manipulations.

One may wonder what the heavy-traffic limits look like for the case of renewal arrivals. To this end, let us assume that the arrival process at Q_i is renewal, with mean interarrival time $1/\lambda_i$ and with variance $Var[A_i]$, for $i = 1, \ldots, N$. Moreover, define

$$\sigma^2 := \sum_{i=1}^{N} \hat{\lambda}_i \left(Var[B_i] + \hat{\rho}_i^2 Var[\hat{A}_i] \right). \tag{33}$$

Note that for the special case of Poisson arrivals we have $Var[\hat{A}_i] = 1/\hat{\lambda}_i^2$, which implies $\sigma^2 = b^{(2)}/b^{(1)}$.

Conjecture 1. *For cyclic polling models with K_i-phase gated service at Q_i and renewal arrivals, we have, for $i = 1, \ldots, N$,*

$$(1-\rho) W_i \to_d \tilde{W}_i \quad (\rho \uparrow 1) \text{ with } \tilde{W}_i =_d U_i \Gamma, \tag{34}$$

where U_i and Γ are independent random variables, and where U_i is uniformly distributed over the interval $[K_i - 1 + \hat{\rho}_i, K_i]$, and Γ is a gamma-distributed random variable with parameters

$$\alpha = \frac{2r\delta}{\sigma^2} + 1, \quad \mu = \frac{2\delta}{\sigma^2} \quad \text{and} \quad \delta = \frac{1}{2} \sum_{j=1}^{N} \hat{\rho}_i ((2K_i - 1) + \hat{\rho}_i), \tag{35}$$

and where σ^2 is defined in (33)*.*

The conjecture can be obtained following the same lines of argumentation as in the derivation of the results in [18] for the case of renewal arrivals with mixtures of exhaustive and gated service at all queues. The following result is an immediate consequence of Conjecture 1.

Conjecture 2 (Moments of the asymptotic delay)**.** *For the case of renewal arrivals, the k-th moment of the asymptotic delay at Q_i is given by the following expression: For $i = 1, \ldots, N$, $k = 1, 2, \ldots$,*

$$E[\tilde{W}_i^k] = E[U_i^k] E[\Gamma^k] = \frac{K_i^{k+1} - (K_i - 1 + \hat{\rho}_i)^{k+1}}{(k+1)(1-\hat{\rho}_i)} \prod_{m=1}^{k} \left[r + \frac{m\sigma^2}{2\delta} \right], \tag{36}$$

where σ^2 and δ are defined in (33) *and* (35)*, respectively.*

We end this section with a number of remarks.

Remark 3. Substituting $k = 1$ in (32) leads to the following expression for the mean scaled delay in heavy traffic: For $i = 1, \ldots, N$,

$$E[\tilde{W}_i] = \frac{2K_i - 1 + \hat{\rho}_i}{2} \left[r + \frac{b^{(2)}/b^{(1)}}{\sum_{j=1}^{N} \hat{\rho}_j (2K_j - 1 + \hat{\rho}_j)} \right]. \tag{37}$$

Equation (37) is remarkable in the sense that the mean asymptotic delay at Q_i in the K_i-phase gated system is proportional to the factor $2K_i - 1 + \hat{\rho}_i$, whereas the results in [24] show that for

$K_i = 1$ the asymptotic mean delay at Q_i is proportional to $1+\hat{\rho}_i$. To provide intuition behind this observation, it is convenient to relate the mean waiting times to the residual cycle times. Defining a cycle time C_i as the time between two successive arrivals of the server to Q_i, it is known that for the case of single-phase gated polling we have the convenient relation, for $K_i = 1$ and $\rho < 1$,

$$E[W_i] = (1+\rho_i)E[RC_i], \tag{38}$$

where RC_i is the residual cycle time. This relation stems from the fact that the expected delay at Q_i consists of two parts (see also [23]): (a) the amount of time until the next visit of the server to Q_i, which takes on average a mean residual cycle time $E[RC_i]$, and (b) the amount of work that arrived during the past cycle time, which has mean value $\rho_i E[RC_i]$; note that $E[RC_i] = E[C_i^2]/2E[C_i]$, with $E[C_i] = r/(1-\rho)$ $(i = 1, \ldots, N)$. On the basis of this relation Groenendijk [12] proposes an approximation for $E[W_i]$ simply by *assuming* that $E[C_i^2]$ is the same for all i, and substituting this into the PCL (4). Numerical results show that the approximation works well for medium and heavily loaded systems.

One may wonder whether a simple cycle-time expression similar to (38) can be obtained for $K_i > 1$. Unfortunately, this is not the case. To this end, note that the mean delay of tagged customer T_i at Q_i can be seen as the sum of the two components (a) and (b) mentioned above, *plus* the $K_i - 1$ (length-biased, see [18]) cycle times during which the customer proceeds along the successive phases $2, 3, \ldots, K_i$: For $i = 1, \ldots, N$, $\rho < 1$,

$$E[W_i] = (1+\rho_i)E\left[RC_i\right] + \sum_{k=2}^{K_i} E\left[C_i^{(k)}\right], \tag{39}$$

where $C_i^{(k)}$ is the duration of a length-biased cycle time in which T_i resides in the phase-k buffer $(k = 2, \ldots, K_i)$. Note that this result can also be directly obtained directly from the cycle-time representation in (99) in Appendix C, and that for $K_i = 1$ the last $K_i - 1$ terms in (39) vanish. In general, for $K_i > 1$ the mean waiting times depend on the correlations between the cycle time at which T_i arrives and the $K_i - 1$ preceding length-biased cycle times $C_i^{(k)}$ $(k = 2, \ldots, K_i)$, whose mean values can not be expressed in closed form. In this context, note that in heavy-traffic, the time-scale decomposition suggested by Coffman et al. [9, 10] implies that the K_i successive cycle-time distributions (properly scaled) are the same for all queues, so that $\lim_{\rho\uparrow 1}\ (1-\rho)E[RC_i] = \lim_{\rho\uparrow 1}\ (1-\rho)E[C_i^{(k)}]$ $(k = 2, \ldots, K_i)$ for all i, so that (39) implies that $E[\tilde{W}_i]$ is proportional to $(1+\hat{\rho}_i) + 2(K_i - 1) = 2K_i - 1 + \hat{\rho}_i$.

Remark 4. Note that for special case $k = 1$ the correctness of (36) was rigorously proven in Van der Mei and Winands [29] using mean value analysis.

Remark 5. It is readily verified that the LST of $\tilde{W}_i$, defined in (34), is given by the expression in (28), but with

$$\alpha = \frac{2r\delta}{\sigma^2}, \quad \mu = \frac{2\delta}{\sigma^2} \quad \text{and} \quad \delta = \frac{1}{2}\sum_{j=1}^{N} \hat{\rho}_i((2K_i - 1) + \hat{\rho}_i), \tag{40}$$

and where σ^2 is defined in (33).

4 Approximation

The results presented in Section 3 suggest the following simple approximations for the moments and the distributions of the waiting times for stable systems: for $\rho < 1$, $i = 1, \ldots, N$, $k = 1, 2, \ldots$,

$$E[W_i^k] \approx E[W_i^k(app)] := \frac{E[\tilde{W}_i^k]}{(1-\rho)^k}, \tag{41}$$

and for $x > 0$,

$$\Pr\{W_i < x\} \approx \Pr\{\tilde{W}_i < x(1-\rho)\}. \tag{42}$$

where closed-form expressions for $E[\tilde{W}_i^k]$ in (41) can be directly obtained from (32) and (36), and where $\Pr\{\tilde{W}_i < x(1-\rho)\}$ in (41) can be calculated by standard one-dimensional inversion of the LST of $\tilde{W}_i$ in (28), for which highly efficient techniques are available (see [1] for details). We have performed numerical experiments to test the accuracy of the approximations in (41) for different values of the load of the system. The relative error of the approximation of the k-th moment of the waiting times at Q_i is defined as follows:

$$\Delta\% := \text{abs}\left(\frac{E[W_i^k(app)] - E[W_i^k]}{E[W_i^k]}\right) \times 100\%. \tag{43}$$

The results of the experiments are outlined below.

Example 1. Consider the model defined by the following parameters: $N = 3$, $K_1 = 2$, $K_2 = 3$, $K_3 = 1$. The ratios of arrival rates at the queues are $1 : 1 : 1$. The service times at queue 1, 2 and 3 are exponentially distributed with mean 3, uniformly distributed over the interval $[1, 3]$, and gamma distributed with mean 1 and variance 2, respectively. The switch-over times from queue 1 to queue 2 and from queue 2 to queue 3 are exponentially distributed with mean 0.1, and the switch-over times from queue 3 to queue 1 are gamma distributed with mean 0.5 and variance 1. Note that the mean total switch-over times per cycle $r = 0.7$ is rather small, compared to the service times. Table 1 shows the "exact" and approximated values of the k-th moments of the waiting times at queue 1 and queue 3, for $k = 1, 3$. The exact values have been obtained from simulations, the approximations are based on (41) and the relative error has been calculated according to (43); confidence intervals are omitted for compactness.

	Queue 1						Queue 3					
	$k=1$			$k=3$			$k=1$			$k=3$		
ρ	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$
0.50	6.42	4.94	30.0	9.77e2	7.77e2	25.7	2.14	2.63	18.6	5.36e1	1.86e2	71.0
0.70	10.7	9.15	16.9	4.53e3	4.10e3	10.2	3.57	4.37	18.3	2.48e2	6.90e2	64.0
0.80	16.1	14.4	11.8	1.53e4	1.45e4	5.52	5.35	6.37	16.1	8.37e2	1.89e3	55.7
0.90	32.1	30.4	5.59	1.22e5	1.20e5	1.67	10.7	12.0	10.8	6.70e3	1.09e4	38.5
0.95	64.2	62.4	2.88	9.78e5	9.70e5	0.82	21.4	22.9	6.55	5.36e4	6.95e4	22.9
0.98	160.5	158.7	1.13	1.53e7	1.52e7	0.66	53.5	55.2	3.08	8.37e5	9.31e5	10.2

Table 1. Exact and approximated values $E\left[W_i^k\right]$ ($i = 1, 3$, $k = 1, 3$) for different values of the load for an asymmetric three-queue model, with $r = 0.7$.

Example 2. To assess the accuracy of the approximations for systems with a larger number of queues, we also consider the seven-queue model with the following parameters: $K_1 = 1$, $K_2 = K_3 = K_4 = K_5 = K_6 = 2$, $K_7 = 4$. The arrival rates are the same for all queues. The service times at queue 1 are gamma distributed with mean 0.5 and variance 4, whereas the service times at all other queues are exponentially distributed with mean 1.5. The switch-over times from queue 1 to queue 2 are uniformly distributed over the interval $[0.05, 0.15]$, and all other switch-over times are exponential with mean 0.25. Note that the mean total switch-over times per cycle equals $r = 1.6$. Table 2 shows the exact and approximated values of the k-th moments of the waiting time at queue 1 and queue 5, for $k = 1, 3$. Again, the exact values have been obtained from simulations, the approximations are based on (41) and the relative error has been calculated according to (43).

To assess the accuracy of the approximations when the switch-over times are large, Table 3 shows the exact and approximated values of the k-th moments of the waiting time at queue 1 and queue

	Queue 1						Queue 5					
	$k=1$			$k=3$			$k=1$			$k=3$		
ρ	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$
0.50	2.63	2.72	3.31	7.64e1	1.64e2	53.4	7.88	6.60	19.4	1.22e3	9.89e2	23.4
0.70	4.38	4.68	6.41	3.54e2	6.31e2	43.9	13.1	11.9	10.1	5.65e3	5.22e3	8.24
0.80	6.56	6.99	6.15	1.19e3	1.83e3	35.0	19.7	18.5	6.49	1.91e4	1.83e4	4.37
0.90	13.1	13.7	4.38	9.55e3	1.21e4	21.1	39.4	38.2	2.87	1.53e5	1.50e5	2.00
0.95	26.3	26.9	2.23	7.64e4	8.63e4	11.5	78.8	77.6	1.42	1.22e6	1.21e6	0.83
0.98	65.6	66.3	1.06	1.19e6	1.25e6	4.80	196.9	195.8	0.56	1.91e7	1.90e7	0.53

Table 2. Exact and approximated values $E\left[W_i^k\right]$ $(i = 1, 5,\ k = 1, 3)$ for different values of the load for an asymmetric seven-queue model ($r = 1.6$).

	Queue 1						Queue 5					
	$k=1$			$k=3$			$k=1$			$k=3$		
ρ	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$
0.50	17.8	18.1	1.66	1.19e4	1.50e4	20.7	53.4	51.9	2.89	1.89e5	1.87e5	1.07
0.70	29.6	30.1	1.66	5.49e4	6.43e4	14.6	88.9	87.5	1.60	8.77e5	8.72e5	0.57
0.80	44.5	45.0	1.11	1.85e5	2.07e5	10.6	133.4	132.0	1.06	2.96e6	2.95e6	0.34
0.90	88.9	89.5	0.67	1.48e6	1.57e6	5.73	266.8	265.4	0.49	2.37e7	2.36e7	0.11
0.95	177.8	178.4	0.34	1.19e7	1.22e7	2.46	533.5	532.2	0.28	1.89e8	1.89e8	0.09
0.98	444.6	445.2	0.01	1.85e8	1.87e8	1.07	1334	1332	0.23	2.96e9	2.96e9	0.01

Table 3. Exact and approximated values $E\left[W_i^k\right]$ $(i = 1, 5,\ k = 1, 3)$ for different values of the load for an asymmetric seven-queue model, with large switch-over times ($r = 16$).

5, for $k = 1, 3$, where the switch-over times are multplied by a factor 10 compared to the models in Table 2.

The results in Tables 1, 2 and 3 reveal a number of observations. First, we observe that the approximations become more accurate when the load is increased, and that the relative error tends to 0 when the system tends to saturate. These observations were expected on the basis of the asymptotic results shown in this paper (Theorem 2). Second, we observe that the accuracy of the approximations tend to degrade for larger values of k. That is, the approximations tend to become less accurate for the higher moments of the delay. This observation can also be explained by the fact that deviations from the limiting waiting-time distribution (for $\rho \uparrow 1$) are "magnified" by taking higher moments. Lastly, we observe that the accuracy of the approximations tends to become better when the switch-over times are large. This observation is in line with similar observations made in [27, 18, 25].

Modified approximation

Expression (37) implies that in the limiting case $\rho \uparrow 1$ it holds that the ratios of the mean waiting times converge to a known limit: For $i, j = 1, \ldots, N$,

$$\lim_{\rho\uparrow 1} \frac{E[W_i]}{E[W_j]} = \frac{2K_i - 1 + \hat{\rho}_i}{2K_j - 1 + \hat{\rho}_j}. \tag{44}$$

This result (44) suggests the following modification to the approximation for the mean waiting times in (41), by combining (44) with the PCL formulated in (4)–(5): For $\rho < 1$, $i = 1, \ldots, N$,

$$E_{mod}[W_i] := \frac{(2K_i - 1 + \rho_i)}{1 - \rho}\ x, \tag{45}$$

where x can be directly obtained by substituting (45) in (4)–(5). Note that both approximations

(i.e., $E_{app}[W_i]$ defined in (41) and $E_{mod}[W_i]$ defined in (45)) are asymptotically exact, satisfying the limiting behavior in Corollary 3 and (37). Recall that the refined approximation generalizes, and follows the same line of argumentation of, the one proposed by Groenendijk [12], which is based on the *assumption* that the mean residual cycle times are the same for all queues, and which was shown to work very well for a wide range of load values.

To assess the accuracy of this modified approximation, Tables 4 and 5 below show the exact (i.e., simulated) and approximated results for the models considered in Tables 2 and 3, respectively. The column indicated as "mod" gives the mean waiting times based on (45), and the relative error is given in the column right next to that.

	Queue 1					Queue 5				
ρ	exact	app	Δ%	mod	Δ%	exact	app	Δ%	mod	Δ%
0.50	2.72	2.63	3.31	2.14	21.3	6.60	7.88	19.4	6.43	2.58
0.70	4.68	4.38	6.41	3.89	16.9	11.9	13.1	10.1	11.7	1.68
0.80	6.99	6.56	6.15	6.08	13.0	18.5	19.7	6.49	18.2	1.62
0.90	13.7	13.1	4.38	12.6	8.03	38.3	39.4	2.87	37.9	1.04
0.95	26.9	26.3	2.23	25.8	4.09	77.7	78.8	1.42	77.3	0.51
0.98	66.3	65.6	1.06	65.2	1.66	195.8	196.9	0.56	195.5	0.15

Table 4. Exact and approximated values $E[W_i]$ $(i = 1, 5)$ for different values of the load for an asymmetric seven-queue model $(r = 1.6)$.

	Queue 1					Queue 5				
ρ	exact	app	Δ%	mod	Δ%	exact	app	Δ%	mod	Δ%
0.50	18.1	17.8	1.66	17.2	4.97	51.9	53.4	2.89	51.7	0.39
0.70	30.1	29.6	1.66	29.1	3.32	87.5	88.9	1.60	87.2	0.34
0.80	45.0	44.5	1.11	43.9	2.44	132.0	133.4	1.06	131.7	0.23
0.90	89.5	88.9	0.67	88.4	1.23	265.5	266.8	0.49	265.1	0.15
0.95	178.4	177.8	0.34	177.3	0.62	532.0	533.5	0.28	531.8	0.04
0.98	444.7	444.6	0.02	444.0	0.16	1331	1334	0.23	1332	0.08

Table 5. Exact and approximated values $E[W_i]$ $(i = 1, 5)$ for different values of the load for an asymmetric seven-queue model $(r = 16)$.

The results in Tables 4 and 5 show that the modified approximation generally does not lead to better results than the approximation defined in (41), indicated as "app". Note that the asymptotic correctness of (45) is confirmed by the results in Tables 4 and 5.

Extension to renewal arrivals

In the previous section we formulated conjectures about the heavy-traffic limits for the case of renewal arrivals, which also led to an approximation for the moments of the waiting times for the stable systems (41). To test the accuracy of the approximations, Table 6 presents the results for the three-queue models considered in Table 1, but where the interarrival times are Erlang-distributed with squared coefficient of variation 0.25.

Similarly, Table 7 shows the results for the case where the interarrival times follow a two-phase hyper-exponential distribution (with balanced means) with squared coefficient of variation 4. The results in Tables 6 and 7 show that the accuracy of the approximations is comparable to the case of Poisson arrivals (Table 1).

	Queue 1						Queue 3					
	$k=1$			$k=3$			$k=1$			$k=3$		
ρ	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$
0.50	4.77	3.33	43.2	3.26e2	2.37e2	37.6	1.59	2.13	25.4	1.79e1	9.69e1	81.5
0.70	7.96	6.23	27.8	1.51e3	1.30e3	16.2	2.65	3.27	19.0	8.27e1	2.77e2	70.1
0.80	11.9	10.1	17.8	5.09e3	4.66e3	9.23	3.98	4.64	14.2	2.79e2	6.87e2	59.4
0.90	23.9	21.9	9.13	4.07e4	3.94e4	3.30	7.96	8.68	8.29	2.23e3	3.68e3	39.4
0.95	47.7	45.8	4.15	3.26e5	3.21e5	1.56	15.9	16.7	4.79	1.79e4	2.30e4	22.2
0.98	119.4	117.3	1.79	5.04e6	5.04e6	0.99	39.8	40.5	1.73	2.79e5	3.07e5	9.12

Table 6. Exact and approximated values $E\left[W_i^k\right]$ $(i = 1, 3,\ k = 1, 3)$ for different values of the load for an asymmetric three-queue model, with Erlang-distributed interarrival times with squared coefficient of variation 0.25.

	Queue 1						Queue 3					
	$k=1$			$k=3$			$k=1$			$k=3$		
ρ	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$	app	exact	$\Delta\%$
0.50	13.0	7.45	74.5	1.07e4	2.61e3	310.0	4.33	3.31	30.8	5.84e2	4.12e2	41.7
0.70	21.7	16.6	30.7	4.93e4	2.53e4	94.9	7.22	6.71	7.60	2.95e3	2.95e3	8.47
0.80	32.5	27.9	16.5	1.67e5	1.15e5	45.2	10.8	11.0	1.82	9.13e3	1.14e4	19.9
0.90	65.0	61.1	6.38	1.33e6	1.15e6	15.7	21.7	23.0	5.65	7.30e4	9.26e4	21.2
0.95	130.0	126.8	2.52	1.07e7	1.00e7	7.00	43.3	46.0	5.87	5.84e5	6.94e5	15.9
0.98	325.1	321.4	1.15	1.67e8	1.64e8	1.83	108.4	112.0	3.21	9.13e6	9.95e6	8.24

Table 7. Exact and approximated values $E\left[W_i^k\right]$ $(i = 1, 3,\ k = 1, 3)$ for different values of the load for an asymmetric three-queue model, with hyper-exponential interarrival times with squared coefficient of variation 4.

5 Optimization

The choice of the control parameter $\underline{K} = (K_1, \ldots, K_N)$ opens up possibilities for optimization. In this section we will use the asymptotic results discussed in Section 4 to develop and evaluate simple yet effective rules for determining the optimal values of $\underline{K}$. To this end, we consider the following optimization problem.

Optimization Problem. *For given weights $c_1, \ldots, c_N \geq 0$, find $\underline{K}^* = (K_1^*, \ldots, K_N^*) \in \mathcal{S}$ that minimizes*

$$C(\underline{K}) = \sum_{i=1}^{N} c_i E[W_i] \tag{46}$$

over all possible values of $\underline{K} \in \mathcal{S} = \{1, 2, \ldots\}^N$.

In the absence of closed-form expressions for $E[W_i]$ for $\rho < 1$, this problem is very hard to solve in general, and $\underline{K}^*$ will generally depend on the loads offered to each of the queues. However, the proper choice of the interleaving levels $\underline{K}^*$ is most critical when the system is heavily loaded. Therefore, in this section we will use the heavy-traffic asymptotic presented in Section 4 to develop approximations for $\underline{K}^*$. In Section 5.1 we present and evaluate the approximations for the case of zero switch-over times, and in Section 5.2 we consider the case of non-zero switch-over times. Section 5.3 ends this section with a discussion.

5.1 Zero switch-over times

In this section we assume that the switch-over times are zero. To start, note that it follows directly by substituting $r = 0$ in (4) that if the ratio c_i/ρ_i is the same for all $i = 1, \ldots, N$, then the cost function $C(\underline{K})$ is the same for all $\underline{K} \in \mathcal{S}$, so that any choice of $\underline{K}$ is optimal. Therefore, throughout this section we will assume that $c_i/\rho_i \neq c_j/\rho_j$ for some $i, j = 1, \ldots, N$. Then without loss of generality, there exists an index p such that

$$\frac{c_1}{\rho_1} \geq \cdots \geq \frac{c_p}{\rho_p} > \frac{c_{p+1}}{\rho_{p+1}} = \cdots = \frac{c_N}{\rho_N}, \text{ and hence, } \frac{c_1}{\hat{\rho}_1} \geq \cdots \geq \frac{c_p}{\hat{\rho}_p} > \frac{c_{p+1}}{\hat{\rho}_{p+1}} = \cdots = \frac{c_N}{\hat{\rho}_N}. \tag{47}$$

The following result shows that under heavy-traffic assumptions, the interleaving levels should be taken either 1 or infinity, and where the "break point" is between p and $p+1$.

Lemma 1 (Optimal policy in the limiting case)**.** *If the switch-over times are zero, then in the limiting case, the setting $K_1^* = \cdots = K_p^* = 1$ and $K_{p+1}^* = \cdots = K_N^* = \infty$ is optimal.*

Proof. Assuming the ordering in (47), it is readily seen that if we denote $x_i := 2K_i - 1 + \hat{\rho}_i > 0$ then the cost function in (46) can then be written as

$$C(\underline{K}) = \frac{c_1 x_1 + \cdots + c_N x_N}{\hat{\rho}_1 x_1 + \cdots + \hat{\rho}_N x_N} B, \quad \text{with} \quad B := \frac{1}{1-\rho} \cdot \frac{b^{(2)}}{2b^{(1)}}. \tag{48}$$

The derivative of (48) with respect to x_i is

$$\frac{\partial C(\underline{K})}{\partial x_i} = \frac{B}{(\hat{\rho}_1 x_1 + \cdots + \hat{\rho}_N x_N)^2} \times \left[x_1(c_i\hat{\rho}_1 - \hat{\rho}_i c_1) + \cdots + x_{i-1}(c_i\hat{\rho}_{i-1} - \hat{\rho}_i c_{i-1}) \right.$$
$$\left. + x_{i+1}(c_i\hat{\rho}_{i+1} - \hat{\rho}_i c_{i+1}) + \cdots + x_N(c_i\hat{\rho}_N - \hat{\rho}_i c_N) \right].$$

Note that the sign of this derivative does not depend on the value of x_i itself. It can easily be verified that $c_i\hat{\rho}_j - \hat{\rho}_i c_j \leq 0$ for $j < i$ and that $c_i\hat{\rho}_j - \hat{\rho}_i c_j \geq 0$ for $j > i$, because of (47). Since this derivative with respect to x_1 is therefore always non-negative, this immediately proves that x_1 should be taken as small as possible, thus $K_1^* = 1$. For x_N this derivative is always non-positive, hence x_N should be taken as large as possible, thus $K_N^* = \infty$. Given these values of K_1^* and K_N^*, the derivatives of (48) with respect to $x_2, \ldots, x_p$ are then always non-negative as well, meaning that $K_2^* = \cdots = K_p^* = 1$. Finally, the derivatives with respect to $x_{p+1}, \ldots, x_{N-1}$ are all non-positive, thus $K_{p+1}^* = \cdots = K_{N-1}^* = \infty$. □

This asymptotic optimum suggests the following heuristic rule for approximating the optimum interleaving level $\underline{K}^*$.

Heuristic rule. *If the switch-over times are zero, then in the case $\rho < 1$ the following heuristic rule applies: if $c_i/\rho_i > \min_{j=1,\ldots,N} c_j/\rho_j$ then $K_i^{(app)} = 1$, and $K_i^{(app)} = K^{(max)}$ otherwise. Here $K^{(max)} \leq \infty$ is an upper bound on the number of phases that can be assigned to queue i.*

Note that Lemma 1 explicity relies on the assumption that $K^{(max)} = \infty$. In case $K^{(max)} < \infty$, a simple rule for the asymptotically optimal interleaving scheme $\underline{K}^*$ cannot be obtained in general. Therefore, to propose a simple rule for the case $K^{(max)} < \infty$ we simply adopt the idea from Lemma 1 that the queues i for which $c_i/\rho_i \leq \min_{j=1,\ldots,N} c_j/\rho_j$ are assigned the maximum interleaving level $K^{(max)}$. The numerical results discussed below show that this leads to highly accurate approximations.

Example 3. Consider the four-queue model defined as follows: the service times are exponentially distributed with mean 3, 1, 2 and 1 respectively and arrivals occur according to a Poisson process in proportion to 1 : 2 : 2 : 3. Switch-over times are zero (see Corollary 2) and $K^{(max)} = 4$. Table 8 shows the approximated and exact optimum interleaving level $\underline{K}^*$ where the cost vectors $\underline{c} = (1, 1, 2, 1)$ and $\underline{c} = (1, 1, 1, 1)$, for different values of the load ρ. The relative error has been calculated similar to (43).

	$\underline{c}=(1,1,1,1)$, $\underline{K}^{(app)}=(1,1,4,1)$				$\underline{c}=(1,1,2,1)$, $\underline{K}^{(app)}=(4,1,1,4)$			
ρ	$\underline{K}^*$	$C(\underline{K}^*)$	$C(\underline{K}^{(app)})$	$\Delta\%$	$\underline{K}^*$	$C(\underline{K}^*)$	$C(\underline{K}^{(app)})$	$\Delta\%$
0.50	$(1,1,4,2)$	6.909	6.944	0.507	$(3,1,1,4)$	8.618	8.645	0.313
0.70	$(2,1,4,1)$	15.59	15.64	0.321	$(4,1,1,3)$	19.55	19.63	0.409
0.90	$(1,1,4,1)$	57.32	57.32	0	$(4,1,1,3)$	72.47	72.58	0.152
0.95	$(1,1,4,1)$	118.7	118.7	0	$(4,1,1,4)$	150.4	150.4	0

Table 8. Approximated and optimal values of the interleaving level $\underline{K}$.

It can be seen that the heuristic rule performs very well in this case: the difference between the approximation and the exact solution is much less than 0.5% even for small values of the load. The general observation is that the approximation becomes more accurate for higher values of the load. This was to be expected because when the load tends to unity, the approximation approaches optimality.

5.2 Non-zero switch-over times

Let us now examine the case of non-zero switch-over times (i.e., $r > 0$). To start, let us assume we have an additional constraint that the number of phases is the same for each queue, i.e., $K_1 = \cdots = K_N = K$. In that case, the optimization function (46) can be rewritten as

$$C(K) = \frac{1}{1-\rho}\left(\frac{\sum_{i=1}^N c_i(L+\hat{\rho}_i)}{\sum_{i=1}^N \hat{\rho}_i(L+\hat{\rho}_i)}B + \frac{1}{2}r\sum_{i=1}^N c_i(L+\hat{\rho}_i)\right),$$

with $L = 2K-1$ and $B = b^{(2)}/2b^{(1)}$. Taking K as a continuous variable, we obtain

$$\frac{\mathrm{d}C(K)}{\mathrm{d}K} = \frac{1}{1-\rho}\left(2B\frac{\sum_{i=1}^N c_i \sum_{i=1}^N \hat{\rho}_i^2 - \sum_{i=1}^N c_i\hat{\rho}_i}{\left(L+\sum_{i=1}^N \hat{\rho}_i^2\right)^2} + r\sum_{i=1}^N c_i\right).$$

Substitution of $A = \sum_{i=1}^N \hat{\rho}_i^2$, $C = \sum_{i=1}^N c_i$ and $D = \sum_{i=1}^N c_i\hat{\rho}_i$ gives

$$\frac{\mathrm{d}C(K)}{\mathrm{d}K} = \frac{1}{1-\rho}\left(2B\frac{CA-D}{(L+A)^2} + rC\right).$$

Now a couple of observations can be made. First, L is discrete-valued and can take the odd values $1, 3, 5, \ldots$. Second, the sign of the term $CA-D$ determines to great extent whether the derivative becomes zero for some L. Third, if L is increased the derivative goes to rC monotonically. These observations lead to the following simple heuristic rule for finding an approximation for the optimal value of K for $\rho < 1$.

Heuristic rule. *If the switch-over times are non-zero, then the following heuristic rule applies: if* $CA - D \geq 0$ *then* $K^{(app)} = 1$, *else if* $2B\frac{CA-D}{(1+A)^2} + rC \geq 0$ *then* $K^{(app)} = 1$, *else* $K^{(app)} = \left[\frac{1}{2}\left(\sqrt{2B\frac{CA-D}{-rC}} - A + 1\right)\right]$.

Note that in the limiting case the approximation lies only in the fact that rounding takes place. To assess the accuracy of this heuristic rule, we have performed extensive numerical experimentation based on simulations. The results are outlined below.

Example 4. Consider the five-queue model defined as follows: the arrival rates at all queues are the same, the service times at queue i are log-normally distributed with $\sigma = 1$ and $\mu = \frac{i}{2}$, $i =$

$1, \ldots, 5$, and the switch-over times between all queues are gamma distributed with mean 0.1 and variance 0.05. Table 9 shows the approximated and exact optimum value of K for $\underline{c} = (0, 0, 0, 0, 1)$ and $\underline{c} = (1, 1, 1, 1, 1)$, and for different values of the load ρ. The relative error has been calculated similar to (43).

	$\underline{c} = (1,1,1,1,1)$, $K^{(app)} = 1$				$\underline{c} = (0,0,0,0,1)$, $K^{(app)} = 2$			
ρ	K^*	$C(K^*)$	$C(K^{(app)})$	$\Delta\%$	K^*	$C(K^*)$	$C(K^{(app)})$	$\Delta\%$
0.50	1	89.14	89.14	0	1	20.32	20.90	2.85
0.70	1	202.4	202.4	0	2	47.40	47.40	0
0.90	1	759.2	759.2	0	2	177.7	177.7	0
0.95	1	1631	1631	0	2	371.3	371.3	0

Table 9. Approximated and optimal values of K.

In the situation where $\underline{c} = (0, 0, 0, 0, 1)$, the heuristic rule is optimal in all cases except for small values of the load. But the difference then is almost negligible. These situations put confidence in the accuracy of the heuristic rule. Note that for the case $\underline{c} = (1, 1, 1, 1, 1)$, in all cases considered the approximation was found to be even identical to the real optimum. Note however that the usefulness of this heuristic rule is a bit limited. Only in extreme cases where the total switch-over time is small, the service times have a very high variability and the costs fulfil exactly the right conditions, the value of K is unequal to 1. Nevertheless, in these cases the heuristic rule performs very well too.

The most general case is the one in which the number of phases in each queue may be different. The optimization problem is here to find $\underline{K}^*$ that minimizes

$$C(\underline{K}) = \frac{1}{1-\rho}\left(\frac{\sum_{i=1}^N c_i x_i}{\sum_{i=1}^N \hat{\rho}_i x_i} B + \frac{1}{2} r \sum_{i=1}^N c_i x_i\right), \tag{49}$$

where $x_i = 2K_i - 1 + \hat{\rho}_i$ and $B = b^{(2)}/2b^{(1)}$. This is a difficult problem to solve and one whose solution involves complex conditions that are not suitable for a simple heuristic rule. Therefore, we will develop a heuristic rule by combining the previous obtained results. If $r = 0$, equation (49) reduces to (48), which was optimized exactly. Here it is optimal to take some values of K_i to be 1 and others to be ∞. On the other hand, if we only look at the last terms of (49) (the terms containing r), then all K_i should be taken 1. Combining this, it seems reasonable to take at least the K_i to be 1 where both agree on. For the remaining queues, we apply the same heuristic rule where all phases were taken the same. The following heuristic rule is then obtained for approximating the optimal interleaving level $\underline{K}^*$ for $\rho < 1$.

Heuristic rule. *Denote by $\mathcal{V}$ the set of queues for which $c_i/\rho_i > \min c_j/\rho_j$, $i, j = 1, \ldots, N$, and make the substitutions*

$$A = \sum_{i=1}^N \hat{\rho}_i^2 + \sum_{i \in \mathcal{V}} \hat{\rho}_i,$$
$$C = \sum_{i \notin \mathcal{V}} c_i,$$
$$D = \sum_{i \notin \mathcal{V}} \hat{\rho}_i \left(\sum_{i=1}^N c_i \hat{\rho}_i + \sum_{i \in \mathcal{V}} c_i\right).$$

If $i \in \mathcal{V}$ then $K_i^{(app)} = 1$. Otherwise, if $CA - D \geq 0$ or $2B\frac{CA-D}{\left(\sum_{i \notin \mathcal{V}} \hat{\rho}_i + A\right)^2} + rC \geq 0$ then $K_i^{(app)} = 1$, else $K_i^{(app)} = \left[\frac{1}{2}\left(\frac{1}{\sum_{i \notin \mathcal{V}} \hat{\rho}_i}\left(\sqrt{2B\frac{CA-D}{-rC}} - A\right) + 1\right)\right]$.

To assess the accuracy of this heuristic rule, we have performed extensive numerical experimentation based on simulations. The results are outlined below.

Example 5. Consider the model defined in Example 3 with the following adjustment: the switch-over times from queue 1 to queue 2 and from queue 3 to queue 4 are exponentially distributed with mean 0.1, and the switch-over times from queue 2 to queue 3 and from queue 4 to queue 1 are gamma distributed with mean 0.5 and variance 1. Table 10 shows the approximated and exact optimum interleaving level $\underline{K}$, for $\underline{c} = (1,1,1,1)$ and $\underline{c} = (3,2,2,7)$, and for different values of the load ρ. The relative error has been calculated according to (43).

	$\underline{c} = (1,1,1,1)$, $\underline{K}^{(app)} = (1,1,1,1)$				$\underline{c} = (3,2,2,7)$, $\underline{K}^{(app)} = (1,1,2,1)$			
ρ	$\underline{K}^*$	$C(\underline{K}^*)$	$C(\underline{K}^{(app)})$	$\Delta\%$	$\underline{K}^*$	$C(\underline{K}^*)$	$C(\underline{K}^{(app)})$	$\Delta\%$
0.50	$(1,1,1,1)$	16.06	16.06	0	$(1,1,1,1)$	56.48	57.43	1.682
0.70	$(1,1,1,1)$	29.64	29.64	0	$(1,1,2,1)$	102.4	102.4	0
0.90	$(1,1,1,1)$	97.49	97.49	0	$(1,1,2,1)$	324.0	324.0	0
0.95	$(1,1,1,1)$	198.7	198.7	0	$(1,1,2,1)$	655.4	655.4	0

Table 10. Approximated and optimal values of the interleaving level $\underline{K}$.

	$\underline{c} = (6,1,3,5)$, $\underline{K}^{(app)} = (1,4,1,1)$				$\underline{c} = (1,10,10,10)$, $\underline{K}^{(app)} = (9,1,1,1)$			
ρ	$\underline{K}^*$	$C(\underline{K}^*)$	$C(\underline{K}^{(app)})$	$\Delta\%$	$\underline{K}^*$	$C(\underline{K}^*)$	$C(\underline{K}^{(app)})$	$\Delta\%$
0.50	$(1,1,1,1)$	60.75	64.99	6.979	$(3,1,1,1)$	117.2	130.8	11.60
0.70	$(1,2,1,1)$	110.5	114.6	3.710	$(4,1,1,1)$	202.4	219.8	8.597
0.90	$(1,3,1,1)$	350.8	352.1	0.371	$(5,1,1,1)$	581.7	603.9	3.816
0.95	$(1,4,1,1)$	704.6	704.6	0	$(6,1,1,1)$	1119	1139	1.787

Table 11. Approximated and optimal values of the interleaving level $\underline{K}$.

Similarly, Table 11 shows the results for $\underline{c} = (6,1,3,5)$ and $\underline{c} = (1,10,10,10)$. The results show that this heuristic rule also performs very well. This can be explained by a couple of factors. Firstly, because the switch-over times are non-zero the number of phases in each queue are not likely to be large. The heuristic rule takes this into account, since in many cases all but one queue are given a parameter of $K_i = 1$, and the parameter of the remaining queue obeys a square root function. Values of, say, 4 or higher are thus very unlikely. Secondly, the most important queues are given priority over the least important ones. The heuristic rule takes this into account since the queues with a high c_i/ρ_i ratio are immediately given a parameter of one. The parameter of the queues with the smallest ratio can potentially be higher.

5.3 Discussion

The results on optimization discussed in Sections 5.1 and 5.2 are based on the asymptotic results derived in Section 3. In this section we discuss a number of observations on the performance of the system for stable systems (i.e., for $\rho < 1$).

Conjecture 3 (Monotonicity of the mean waiting times)**.** *For $\rho < 1$, it holds that if K_i is increased then $E[W_i]$ is increased, while $E[W_j]$ is decreased for all $j \neq i$.*

Note that the conjecture is supported by numerous numerical experiments, and moreover, is in line with similar observations made for cyclic polling models with K_i-limited [6] and for Bernoulli service disciplines [3]. Note also that it is easily verified from (37) that the conjecture is asymptotically correct for $\rho \uparrow 1$. The following result is an immediate consequence of Conjecture 3.

Corollary 4. *For $\rho < 1$ and $r \geq 0$, it holds that*

$$\text{if } \frac{c_i}{\rho_i} = \max\left\{\frac{c_1}{\rho_1}, \ldots, \frac{c_N}{\rho_N}\right\}, \text{ then } K_i^* = 1. \tag{50}$$

Proof: The result follows directly from the following inequalities:

$$\begin{aligned}
\frac{\Delta}{\Delta k_i}\sum_{j=1}^{N} c_j E[W_j] &= \frac{c_i}{\rho_i}\frac{\Delta}{\Delta k_i}\rho_i E[W_i] + \sum_{j\neq i}\frac{c_j}{\rho_j}\frac{\Delta}{\Delta k_i}\rho_j E[W_j] &(51)\\
&\leq \frac{c_i}{\rho_i}\frac{\Delta}{\Delta k_i}\rho_i E[W_i] + \frac{c_i}{\rho_i}\sum_{j\neq i}\frac{\Delta}{\Delta k_i}\rho_j E[W_j] &(52)\\
&= \frac{c_i}{\rho_i}\frac{\Delta}{\Delta k_i}\sum_{j=1}^{N}\rho_j E[W_j] \leq 0. &(53)
\end{aligned}$$

Here, the first inequality follows from the assumption (47), and the second inequality follows from (4)–(5).

Corollary 5. *For $\rho < 1$ and $r = 0$, it holds that*

$$\text{if } \frac{c_i}{\rho_i} = \min\left\{\frac{c_1}{\rho_1}, \ldots, \frac{c_N}{\rho_N}\right\}, \text{ then } K_i^* = \infty. \tag{54}$$

Proof: The result follows from the following inequalities:

$$\begin{aligned}
\frac{\Delta}{\Delta k_i}\sum_{j=1}^{N} c_j E[W_j] &= \frac{c_i}{\rho_i}\frac{\Delta}{\Delta k_i}\rho_i E[W_i] + \sum_{j\neq i}\frac{c_j}{\rho_j}\frac{\Delta}{\Delta k_i}\rho_j E[W_j] &(55)\\
&\geq \frac{c_i}{\rho_i}\frac{\Delta}{\Delta k_i}\rho_i E[W_i] + \frac{c_i}{\rho_i}\sum_{j\neq i}\frac{\Delta}{\Delta k_i}\rho_j E[W_j] &(56)\\
&= \frac{c_i}{\rho_i}\frac{\Delta}{\Delta k_i}\sum_{j=1}^{N}\rho_j E[W_j] = 0, &(57)
\end{aligned}$$

noting that the last equality follows from (4)–(5) by taking $r = 0$.

Remark 6. It is hard to make general statements about the sensitivity of the waiting-time performance of the system as a function of the interleaving levels $\underline{K} = (K_1, \ldots, K_N)$ for $\rho < 1$. The PCL in (4)–(5) reveals that the summation $\sum_{i=1}^{N}\rho_i E[W_i]$, representing the total amount of waiting work in the system, is primarily sensitive to the choice of $\underline{K}$ when the system is heavily loaded and the switch-over times are large. Extensive numerical experiments show similar observations regarding the individual waiting times.

Remark 7. Note that strictly speaking the notion of $K_i = \infty$ (e.g., in Lemma 1 and Corollary 5) is not well defined, and that in practice this means setting K_i to a large but finite value.

Remark 8. Let us reconsider the optimization problem (46) for the case of zero switch-over times (i.e., $r = 0$). Loosely speaking, Corollary 4 states that for stable systems (i.e., $\rho < 1$) the queue(s) with the highest c_i/ρ_i-ratio should get the best available service (i.e., $K_i^* = 1$), whereas the queue(s) with the lowest c_i/ρ_i-ratio should get the worst available service (i.e., $K_i^* = \infty$). Note that this leaves open the possibility that $1 < K_i^* < \infty$ for queues for which the c_i/ρ_i-ratio is in between the minimum and the maximum c_i/ρ_i-ratio. In this context, note that Lemma 1 implies that in the limiting case $\rho \uparrow 1$, K_i^*-values between 1 and ∞ do not occur. More precisely, Lemma 1 implies that $K_i^* = 1$ or $K_i^* = \infty$ for all i, and that $K_i^* = \infty$ if and only if $c_i/\rho_i = \min\{c_i/\rho_i\}$. Apparently, the possibility of having finite K_i^*-values larger than 1 vanishes when $\rho \uparrow 1$.

Remark 9. In the polling literature, cost-optimization problems similar to (46) have been studied. For example, for cyclic polling models, minimization of a cost function of the form $T(\underline{c}) := \sum_{i=1}^{N} c_i E[W_i]$ with respect to the choice of the service policies has been studied by Boxma et al. [5] for K_i-limited service and by Blanc and Van der Mei [3] for Bernoulli service policies. For these models, conjectures similar to Conjecture 3 have been formulated without rigorous proofs. As an alternative, Boxma et al. [7] study the problem of minimizing $T(\underline{c})$ by choosing (near-)optimal periodic visit orders, and derive elegant square-root rules for the relative visit frequencies. For the multi-phase gated model under consideration, minimization of $T(\underline{c})$ with respect to the visit order, following the lines of argumentation of [7] and deriving heavy-traffic limits similar to those in [19], addresses an interesting direction of research.

Acknowledgment: The authors wish to thank the anonymous referees for their useful comments, which have led to a significant improvement of the readability of the paper.

References

[1] Abate, J. and Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. Queueing Systems 10, 5–88.

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

[3] Blanc, J.P.C. and Van der Mei, R.D. (1995). Optimization of polling systems with Bernoulli schedules. Performance Evaluation 2, 139-158.

[4] Borst, S.C. and Boxma, O.J. (1997). Polling models with and without switchover times. Operations Research 45, 536-543.

[5] Boxma, O.J. and Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. Journal of Applied Probability 24, 949-964.

[6] Borst, S.C., Boxma, O.J. and Levy, H. (1995). The use of service limits for efficient operation of multistation single-medium communication systems IEEE/ACM Transactions on Networking 3, 602-612.

[7] Boxma, O.J., Levy, H. and Weststrate, J.A. (1993) Efficient visit frequencies for polling tables: minimization of waiting cost. Queueing Systems 9, 133-162.

[8] Choudhury, G.L. and Whitt, W. (1996). Computing transient and steady state distributions in polling models by numerical transform inversion. Performance Evaluation 25, 267-292.

[9] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1995). Polling systems with zero switchover times: a heavy-traffic principle. Annals Applied Probability 5, 681-719.

[10] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1998). Polling systems in heavy-traffic: a Bessel process limit. Mathematics of Operations Research 23, 257-304.

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

[12] Groenendijk, W.P. (1989). Waiting-time approximations for cyclic-service systems with mixed service strategies. Proceedings ITC 12, 1434-1441.

[13] Konheim, A.G., Levy, H. and Srinivasan, M.M. (1994). Descendant set: an efficient approach for the analysis of polling systems. IEEE Transactions on Communications 42, 1245–1253.

[14] Kramer, G., Muckerjee, B. and Pesavento, G. (2001). Ethernet PON: design and analysis of an optical access network. Photonic Network Communications 3, 307–319.

[15] Kramer, G., Muckerjee, B. and Pesavento, G. (2002). Interleaved polling with adaptive cycle time (IPACT): a dynamic bandwidth allocation scheme in an optical access network. Photonic Network Communications 4, 89–107.

[16] Kramer, G., Muckerjee, B. and Pesavento, G. (2002). Supporting differentiated classes of services in Ethernet passive optical networks. Journal of Optical Netwworks 1, 280–290.

[17] Olsen, T.L. (2001). Limit theorems for polling models with increasing setups. Probability and Engineering of Information Systems 15, 35-55.

[18] Olsen, T.L. and Van der Mei, R.D. (2005). Polling systems with periodic server routing in heavy traffic - renewal arrivals. Operations Research Letters 33, 17-25.

[19] Olsen, T.L. and Van der Mei, R.D. (2003). Polling systems with periodic server routeing in heavy-traffic: distribution of the delay. Journal of Applied Probability 40, 305-326.

[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] Quine, M.P. (1972). The multitype Galton-Watson process with ρ near 1. Advances in Applied Probability 4, 429-452.

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

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

[24] Van der Mei, R.D. (1999). Distribution of the delay in polling systems in heavy traffic. Performance Evaluation 31, 163-182.

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

[26] Van der Mei, R.D. (2007). Towards a unifying theory on branching-type polling models in heavy traffic. Queueing Systems 57, 29-46.

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

[28] Van der Mei, R.D. and Resing, J.A.C. (2008). Polling models with two-phase gated service: heavy-traffic results for the waiting-time distributions. Probability in the Engineering and Informational Sciences 22, 623-651.

[29] 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.

[30] Van der Mei, R.D. and Winands, E.M.M. (2008). A note on polling models with renewal arrivals and nonzero switch-over times. Operations Research Letters 36, 500-505.

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

Appendix A: Multi-type branching processes with immigration in each state

For completeness, in this appendix we briefly describe general MTBPs with immigration in each state, and introduce notation useful for further reference. We refer to [2] for more details on MTBPs. 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 $|z_k| \leq 1$ $(k = 1, \ldots, M), i = 1, \ldots, M$,

$$f^{(i)}(\underline{z}) = \sum_{j_1, \ldots, j_M \geq 0} p^{(i)}(j_1, \ldots, j_M) z_1^{j_1} \cdots z_M^{j_M}, \tag{58}$$

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 defined as follows, for $|z_k| \leq 1$ $(k = 1, \ldots, M)$,

$$g(\underline{z}) = \sum_{j_1, \ldots, j_M \geq 0} q(j_1, \ldots, j_M) z_1^{j_1} \cdots z_M^{j_M}, \tag{59}$$

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

$$\underline{g} := (g_1, \ldots, g_M), \text{ where } g_i := \frac{\partial g(\underline{z})}{\partial z_i}|_{\underline{z}=\underline{1}} \quad (i = 1, \ldots, M), \tag{60}$$

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

$$\mathbf{M} = (m_{i,j}), \quad \text{with } m_{i,j} := \frac{\partial f^{(i)}(\underline{z})}{\partial z_j}|_{\underline{z}=\underline{1}} \quad (i, j = 1, \ldots, M). \tag{61}$$

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

$$\mathbf{K}^{(i)} = \left(k_{j,k}^{(i)}\right), \quad \text{with } k_{j,k}^{(i)} := \frac{\partial^2 f^{(i)}(\underline{z})}{\partial z_j \partial z_k}|_{\underline{z}=\underline{1}} \quad (i, j, k = 1, \ldots, M). \tag{62}$$

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., [2]), normalized such that

$$\underline{v}^\top \underline{1} = \underline{v}^\top \underline{w} = 1. \tag{63}$$

The following conditions are necessary and sufficient conditions for the ergodicity of the process $\mathbf{Z}$ (cf. [22]): $\xi < 1$ and

$$\sum_{j_1 + \cdots + j_M > 0} q(j_1, \ldots, j_M) \log(j_1 + \cdots + j_M) < \infty. \tag{64}$$

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

$$\pi_0(\xi) := 0, \quad \text{and} \quad \pi_n(\xi) := \sum_{r=1}^{n} \xi^{r-2}, \quad n = 1, 2, \ldots. \tag{65}$$

Theorem A.1
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

$$\frac{1}{\pi_n(\xi)} \underline{Z}_n \to A \cdot \hat{\underline{v}} \cdot \Gamma \cdot (\alpha, 1) \quad (\xi, n) \to (1, \infty), \tag{66}$$

in the sense that for all $\epsilon > 0$ there exist $\delta > 0$ and N such that if $|1 - \xi| < \delta$ then for all $n > N$ it holds that

$$\sup_{\underline{x} \in \mathcal{R}^M} \left| \text{Prob} \left\{ \frac{1}{\pi_n(\xi)} \underline{Z}_n \leq \underline{x} \right\} - \text{Prob} \left\{ A \cdot \Gamma(\alpha, 1) \cdot \hat{\underline{v}} \leq \underline{x} \right\} \right| < \epsilon, \tag{67}$$

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

$$\alpha := \frac{1}{A} \hat{\underline{g}}^\top \hat{\underline{w}} = \frac{1}{A} \sum_{i=1}^{M} \hat{g}_i \hat{w}_i, \text{ with } A := \sum_{i=1}^{M} \hat{v}_i \left(\hat{\underline{w}}^\top \hat{\mathbf{K}}^{(i)} \hat{\underline{w}} \right) > 0. \tag{68}$$

Proof: See [21, 26]. □

Appendix B: The Descendant Set Approach for multi-phase gated service

In this appendix we formulate the use of the DSA for the model under consideration. It is a rather straightforward extension of the DSA for the two-phase gated model discussed in [28]. Customers 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 K_1-dimensional stochastic vector

$$\underline{X}_1(P^*) := \left(X_1^{(1)}(P^*), \ldots, X_1^{(K_1)}(P^*) \right), \tag{69}$$

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 $(k = 1, \ldots, K_1)$. P^* is referred to as the *reference point*.

The main ideas are the observations that (a) each of the customers present at Q_1 at the reference point P^* (at either phase) belongs to the descendant set of exactly one originator, and (b) 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 all K_1 phases at P^*. Summing up these numbers over all past originators yields $\underline{X}_1(P^*)$, and hence $\underline{X}_1$, 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 $\underline{X}_1(P^*)$ 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 for $i = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$\underline{A}_{i,c} := \left(A_{i,c}^{(1)}, \ldots, A_{i,c}^{(K_1)}\right), \quad \text{and its PGF} \quad A_{i,c}^*(z_1, \ldots, z_{K_1}) := E\left[z_1^{A_{i,c}^{(1)}} \cdots z_{K_1}^{A_{i,c}^{(K_1)}}\right], \tag{70}$$

where $A_{i,c}^{(k)}$ is the number of type-1 descendants it has at phase k at P^* ($k = 1, \ldots, K_1$). In this way, the K_1-dimensional random variable $\underline{A}_{i,c}$ can be viewed as the contribution of $T_{i,c}$ to $\underline{X}_1(P^*)$.

To express the distribution of $\underline{X}_1$ 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, for $i = 1, \ldots, N$, $c = 0, 1, \ldots$, denote

$$\underline{S}_{i,c} := \left(S_{i,c}^{(1)}, \ldots, S_{i,c}^{(K_1)}\right), \quad \text{and its PGF} \quad S_{i,c}^*(z_1, \ldots, z_{K_1}) := E\left[z_1^{S_{i,c}^{(1)}} \cdots z_{K_1}^{S_{i,c}^{(K_1)}}\right], \tag{71}$$

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). In this way, $\underline{S}_{i,c}$ can be seen as the (joint) contribution of $R_{i,c}$ to $\underline{X}_1(P^*)$. It is readily verified that we can write

$$\underline{X}_1(P^*) = \sum_{i=1}^{N} \sum_{c=0}^{\infty} \underline{S}_{i,c}. \tag{72}$$

Note that for $k, l = 1, \ldots, K_1$, the random variables $S_{i,c}^{(k)}$ and $S_{i',c'}^{(l)}$ are generally dependent if $(i,c) = (i',c')$ but independent otherwise. Hence we can write, for $|z_1|, \ldots, |z_{K_1}| \leq 1$,

$$X_1^*(z_1, \ldots, z_{K_1}) = \prod_{i=1}^{N} \prod_{c=0}^{\infty} S_{i,c}^*(z_1, \ldots, z_{K_1}). \tag{73}$$

Because $\underline{S}_{i,c}$ is the total joint contribution to $\underline{X}_1(P^*)$ of all (original) customers that arrive during $R_{i,c}$, the distribution of $\underline{S}_{i,c}$ can be expressed in terms of the distributions of the descendant set variables $\underline{A}_{i,c}$ as follows: For $i = 1, \ldots, N$, $c = 0, 1, \ldots$, and $|z_1|, \ldots, |z_{K_1}| \leq 1$,

$$\begin{aligned} &S_{i,c}^*(z_1, \ldots, z_{K_1}) = \\ &R_i^*\left(\sum_{j=i+1}^{N} \left[\lambda_j - \lambda_j A_{j,c}^*(z_1, \ldots, z_{K_1})\right] + \sum_{j=1}^{i} \left[\lambda_j - \lambda_j A_{j,c-1}^*(z_1, \ldots, z_{K_1})\right]\right). \end{aligned} \tag{74}$$

Next, 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 cycle $c - K_1 + 1$, because it takes $K_1 - 1$ cycles to proceed along the phases $1 \to 2 \to \cdots \to K_1$. This leads to the following relation: For $i = 1, \ldots, N$, $c = 0, 1, \ldots$, and $|z_1|, \ldots, |z_{K_1}| \leq 1$,

$$\begin{aligned} &A_{i,c}^*(z_1, \ldots, z_{K_1}) = \\ &B_i^*\left(\sum_{j=i+1}^{N} \left[\lambda_j - \lambda_j A_{j,c-K_1+1}^*(z_1, \ldots, z_{K_1})\right] + \sum_{j=1}^{i} \left[\lambda_j - \lambda_j A_{j,c-K_1}^*(z_1, \ldots, z_{K_1})\right]\right), \end{aligned} \tag{75}$$

supplemented with the basis for the recursion: for $i = 1, \ldots, N$, $k = 1, \ldots, K_1$, $|z_1|, \ldots, |z_{K_1}| \leq 1$,

$$A^*_{i,-k}(z_1, \ldots, z_{K_1}) = \begin{cases} z_k & \text{if } i = 1, \\ 1 & \text{if } i > 1. \end{cases} \tag{76}$$

In this way, relations (72)-(76) give a complete, characterization of the distribution of $\underline{X}_1$. Similarly, recursive relations to calculate the (cross-)moments of $\underline{X}_1$ can be readily obtained from those equations.

Relation (75) leads to recursive relations for the moments of the DS variables $A^{(k)}_{i,c}$. Here, we will work out the details of the DSA for calculation the first moment of the DS variables; they are also needed for the proof of Lemma C.4 below. Define for $i = 1, \ldots, N$, $c = -K_1, -K_1+1, \ldots, 0, 1, \ldots$, and $k = 1, \ldots, K_1$,

$$\alpha^{(k)}_{i,c} := E\left[A^{(k)}_{i,c}\right] = \left[\frac{\partial}{\partial z_k} A^*_{i,c}(\underline{z})\right]_{\underline{z}=\underline{1}}. \tag{77}$$

Then (75)-(76) are easily seen to lead to the following recursive scheme: For $i = 1, \ldots, N$, $c = 0, 1, \ldots$, and $k = 1, \ldots, K_1$,

$$\alpha^{(k)}_{i,c} = b^{(1)}_i \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(k)}_{j,c-K_1+1} + \sum_{j=1}^{i} \lambda_j \alpha^{(k)}_{j,c-K_1}\right], \tag{78}$$

supplemented with the following basis for the recursion, for $i = 1, \ldots, N$, $k, l = 1, \ldots, K_1$,

$$\alpha^{(l)}_{i,-k} := \left[\frac{\partial}{\partial z_k} A^*_{i,-k}(\underline{z})\right]_{\underline{z}=\underline{1}} = \begin{cases} 1 & \text{if } i = 1 \text{ and } k = l, \\ 0 & \text{otherwise.} \end{cases} \tag{79}$$

Moreover, recalling that $X_1 := X^{(1)}_1 + \cdots + X^{(K_1)}_1$, it follows directly from (73)-(74) that

$$E[X_1] = \sum_{k=1}^{K_1} E\left[X^{(k)}_1\right], \quad \text{where} \quad E\left[X^{(k)}_1\right] = \sum_{i=1}^{N} \sum_{c=0}^{\infty} E\left[S^{(k)}_{i,c}\right] \quad (k = 1, \ldots, K_1), \tag{80}$$

with

$$E\left[S^{(k)}_{i,c}\right] = r^{(1)}_i \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(k)}_{j,c-K_1+1} + \sum_{j=1}^{i} \lambda_j \alpha^{(k)}_{j,c-K_1}\right]. \tag{81}$$

The calculation of the higher moments of X_1 requires more effort because of the dependence between the random variables $(X^{(1)}_1, \ldots, X^{(K_1)}_1)$, but is methodologically straightforward but notationally cumbersome. Working out the details is beyond the scope of this paper.

Appendix C: Stepwise derivation of Theorem 2

In this section we use the MTBP structure and the DSA discussed in Appendix A and Appendix B, respectively, to transform Theorem A.1 into Theorem 2, which gives an expression for the limiting distribution for $\underline{X}$ as ρ goes to 1. We follow the step-wise approach proposed in [26]. Similar to the derivation of the result for the case of two-phase gated service at all queues [28], we proceed along the following steps:

Step 1: Derive an expression for the mean offspring matrix $\mathbf{M}$ for the polling model under consideration (Lemma C.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 C.2)
Step 3: Derive an expression for the mean immigration vector $\underline{g}$, evaluated at $\rho = 1$ (Lemma C.3).
Step 4: Derive an expression for limiting behavior of $\xi(\rho)$ considered as a function of ρ, as ρ goes to 1 (Lemma C.4).
Step 5: Derive an expression for A, evaluated at $\rho = 1$ (Lemma C.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.

Lemma C.1 (Mean matrix)
The mean offspring matrix $\mathbf{M}$*, defined in* (21)–(24)*, is given by*

$$\mathbf{M} = \mathbf{M_1} \cdots \mathbf{M}_N, \text{ where } \mathbf{M}_k := \begin{pmatrix} \mathbf{T}_{1,1}^{(k)} & \cdots & \mathbf{T}_{1,N}^{(k)} \\ \vdots & \vdots & \vdots \\ \mathbf{T}_{N,1}^{(k)} & \cdots & \mathbf{T}_{N,N}^{(k)} \end{pmatrix} \quad (k = 1, \ldots, N), \tag{82}$$

where for $i, j = 1 \ldots, N$, $\mathbf{T}_{i,j}^{(k)}$ *is a* K_i*-by-*K_j *block matrix with entries* $\{t_{i,j}^{(k)}(m,n),\ m = 1, \ldots, K_i, n = 1, \ldots, K_j\}$*, defined as follows: For* $i \neq k$*,*

$$t_{i,i}^{(k)}(m,n) = I_{\{m=n\}}, \qquad t_{i,j}^{(k)}(m,n) = 0 \text{ for } i \neq j. \tag{83}$$

Moreover, for $k \neq j$*,*

$$t_{k,j}^{(k)}(m,n) = \begin{cases} b_k \lambda_j & \text{if } (m,n) = (K_k, 1), \\ 0 & \text{otherwise}, \end{cases} \tag{84}$$

and

$$t_{k,k}^{(k)}(m,n) = \begin{cases} 1 & \text{if } n = m+1, \\ b_k \lambda_k & \text{if } (m,n) = (K_k, 1), \\ 0 & \text{otherwise}. \end{cases} \tag{85}$$

Proof: The result can be obtained by taking the partial derivatives of the mean offspring function defined in Theorem 1. □

The following result gives the left and right eigenvectors of the mean matrix $\mathbf{M}$, evaluated at $\rho = 1$, and normalized according to (63).

Lemma C.2 (Eigenvectors of mean matrix M at $\rho = 1$)
The normalized right eigenvector of the mean matrix $\hat{\mathbf{M}}$ *is given by* $\underline{\hat{w}} := |\underline{y}|^{-1}\underline{y}$*, with*

$$\underline{y} := \left(b_1^{(1)}, \ldots, b_1^{(1)}, \ldots, b_N^{(1)}, \ldots, b_N^{(1)}\right), \text{ and } |\underline{y}| := \sum_{j=1}^{N} \sum_{k=1}^{K_j} y_j^{(k)} = \sum_{j=1}^{N} K_j b_j^{(1)}. \tag{86}$$

Similarly, the normalized left eigenvector of $\hat{\mathbf{M}}$ *is given by* $\underline{\hat{v}} := \frac{|\underline{y}|}{\delta}\underline{\hat{u}}$*, where*

$$u_j^{(1)} := \lambda_j(\rho_j + \cdots + \rho_N), \quad u_j^{(k)} := \lambda_j \ \ (k = 2, \ldots, K_j), \quad \text{for } j = 1 \ldots, N, \tag{87}$$

and where

$$\delta := \underline{\hat{u}}^\top \underline{y} = \sum_{i=1}^{N} \sum_{j=i}^{N} \hat{\rho}_i \hat{\rho}_j + \sum_{i=1}^{N} (K_i - 1)\hat{\rho}_i I_{\{K_i > 1\}} = \frac{1}{2} \sum_{i=1}^{N} \hat{\rho}_i((2K_i - 1) + \hat{\rho}_i). \tag{88}$$

Proof: It is readily verified from (83)–(85) that $\hat{\mathbf{M}}_k \underline{\hat{w}} = \underline{\hat{w}}$ for $k = 1, \ldots, N$, which immediately

implies $\mathbf{M}\underline{\hat{w}} = \mathbf{M}_1 \cdots \mathbf{M}_N \underline{\hat{w}} = \underline{\hat{w}}$, so that $\underline{\hat{w}}$ indeed is a right eigenvector of $\hat{\mathbf{M}}$. Similar arguments can be used to show that $\hat{\mathbf{M}}^\top \underline{\hat{v}} = \underline{\hat{v}}$, which implies that $\underline{\hat{v}}$ is a left eigenvector of $\hat{\mathbf{M}}$. $\square$

We define the mean immigration vector $\underline{g} := \left(g_1^{(1)}, \ldots, g_1^{(K_1)}, \ldots, g_N^{(1)}, \ldots, g_N^{(K_N)}\right)$ by

$$g_k^{(l)} := \left| \frac{\partial}{\partial s_k^{(l)}} g(\underline{s}) \right|_{\underline{s}=\underline{1}} \qquad (k = 1, \ldots, N,\ l = 1, \ldots, K_k). \tag{89}$$

Lemma C.3
The mean immigration vector $\underline{g}$ satisfies the following equations: For $k = 1, \ldots, N$, $l = 1 \ldots, K_k$,

$$g_k^{(l)} = \sum_{i=1}^{N} r_i \left[\sum_{j=1}^{i} \lambda_j I_{\{l=1,k=j\}} + \sum_{j=i+1}^{N} \lambda_j m_{j,k}^{(1,l)} \right], \tag{90}$$

and moreover,

$$\hat{\underline{g}}^\top \underline{\hat{w}} = |\underline{y}|^{-1} r, \ \textit{with} \ \ |\underline{y}| = \sum_{j=1}^{N} K_j b_j^{(1)}. \tag{91}$$

Proof: Equation (90) can be directly obtained from (18) by differentiating once with respect to $s_k^{(l)}$ and substituting $\underline{s} = (1, \ldots, 1)$. Next, (91) follows directly from the following sequence of relations:

$$\begin{aligned}
\hat{\underline{g}}^\top \underline{\hat{w}} &:= \sum_{k=1}^{N} \sum_{l=1}^{K_k} \hat{g}_k^{(l)} \hat{w}_k^{(l)} = |\underline{y}|^{-1} \sum_{k=1}^{N} b_k^{(1)} \sum_{l=1}^{K_k} \sum_{i=1}^{N} r_i \left[\sum_{j=1}^{i} \hat{\lambda}_i I_{\{l=1,k=j\}} + \sum_{j=i+1}^{N} \hat{\lambda}_j m_{j,k}^{(1,l)} \right] && (92) \\
&= |\underline{y}|^{-1} \sum_{i=1}^{N} r_i \left[\sum_{j=1}^{i} \hat{\lambda}_j b_j^{(1)} + \sum_{j=i+1}^{N} \hat{\lambda}_j \sum_{k=1}^{N} \sum_{l=1}^{K_k} m_{j,k}^{(1,l)} b_k^{(1)} \right] && (93) \\
&= |\underline{y}|^{-1} \sum_{i=1}^{N} r_i \left[\sum_{j=1}^{i} \hat{\lambda}_j b_j^{(1)} + \sum_{j=i+1}^{N} \hat{\lambda}_j b_j^{(1)} \right] = |\underline{y}|^{-1} \hat{\rho} r = |\underline{y}|^{-1} r. && (94)
\end{aligned}$$

The first equality in (92) follows from the definitions of $g_j^{(k)}$ and $\hat{w}_j^{(k)}$, standard algebraic manipulations and by noting that for $j = 1, \ldots, N$ it holds that $\sum_{k=1}^{N} \sum_{l=1}^{K_k} m_{j,k}^{(1,l)} b_k^{(1)} = b_j^{(1)}$, which follows from the fact that $\underline{y}$ is a right eigenvector of $\hat{\mathbf{M}}$, as shown in Lemma C.2. This completes the proof of the result. $\square$

Lemma C.4
The Frobenius eigenvalue $\xi = \xi(\rho)$ of $\mathbf{M}$ 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*

$$\xi'(1) = \lim_{\rho \uparrow 1} \frac{1 - \xi(\rho)}{1 - \rho} = \frac{1}{\delta}, \ \textit{with} \ \delta = \frac{1}{2} \sum_{i=1}^{N} \hat{\rho}_i ((2K_i - 1) + \hat{\rho}_i). \tag{95}$$

Proof: The proof proceeds along similar lines as the one for $K_i = 2$ $(i = 1, \ldots, N)$ in [28]. $\square$

Lemma C.5
For the multi-phase gated polling model,

$$A = |\underline{y}|^{-1}\delta^{-1}\frac{b^{(2)}}{2b^{(1)}}. \tag{96}$$

Proof: The scaling parameter A, defined in (68), result can be directly obtained from Theorem 1 and the definitions in (61)–(63), following rather tedious calculations. Alternatively, A follows directly from Theorem 2 (which does not require the scaling parameter A to be known explicitly), and (12). To this end, note that taking the mean value of the first entry in (25) and combining this with Lemma C.2 implies that $\hat{\lambda}_1 r = \delta A\hat{v}_1 \cdot \alpha = \delta A \cdot \frac{|y|}{\delta}\hat{\lambda}_1 \cdot 2r\delta\frac{b^{(1)}}{b^{(2)}}$, where δ is defined in (95). This immediately implies (96). $\square$

Proof of Theorem 2: Without loss of generality, we assume $i = 1$. In the sequel, it will be convenient to relate the waiting-time and queue-length distributions at polling instants at Q_1 to the joint distribution of K_1 *successive cycle times*. Let a given polling instant P at Q_1 mark the end of a cycle time with duration $C_1^{(1)}$, let the duration of the K_1 *preceding* cycle times be $C_1^{(2)}, \ldots, C_1^{(K_1)}$, and denote the joint LST of $\left(C_1^{(1)}, \ldots, C_1^{(K_1)}\right)$ by:

$$C_1^*(s_1, \ldots, s_{K_1}) := E\left[e^{-s_1 C_1^{(1)} - \cdots - s_{K_1} C_1^{(K_1)}}\right]. \tag{97}$$

Recall from Appendix B that $X_1^*(z_1, \ldots, z_{K_1})$ is the joint PGF of the numbers of type-1 customers at all K_1 phases at Q_1 at an arbitrary polling instant at Q_1. Then the population of customers present at Q_1 at phase k at polling instant P consists exactly of those customers that arrived during the cycle time $C_1^{(k)}$, for $k = 1, \ldots, K_1$. Standard GF manipulations then immediately imply that

$$X_1^*(z_1, \ldots, z_{K_1}) = C_1^*\left(\lambda_1(1 - z_1), \ldots \lambda_1(1 - z_{K_1})\right). \tag{98}$$

Using (98), equation (10) can be reformulated in the following convenient form: For $Re(s) > 0$,

$$W_1^*(s) = \frac{(1-\rho_1)s}{s - \lambda_1(1 - B_1^*(s))} \cdot \frac{C_1^*(s, \ldots, s, \lambda_1(1 - B_1^*(s))) - C_1^*(s, \ldots, s, s)}{s(1-\rho_1)r/(1-\rho)}. \tag{99}$$

Now, combining Theorem A.1 with Lemmas C.2–C.5, and by taking the proper components of the vector $\hat{v}$ defined in Lemma 2, it follows that

$$(1-\rho)\left(X_1^{(1)}, \ldots, X_1^{(K_1)}\right) \to_d \frac{1}{2\delta} \cdot \frac{b^{(2)}}{b^{(1)}}\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_1\right)\Gamma(\alpha, 1) \quad (\rho \uparrow 1), \tag{100}$$

where α and δ are defined in (26) and (95), respectively (see Remark 3.3 in [28] for the details on the convergence). Then, using (98) and similar arguments as those discussed in [28], equation (100) can be expressed in terms of cycle times as

$$(1-\rho)\left(C_1^{(1)}, \ldots, C_1^{(K_1)}\right) \to_d \frac{1}{2\delta} \cdot \frac{b^{(2)}}{b^{(1)}}(1, 1, \ldots, 1)\,\Gamma(\alpha, 1) \quad (\rho \uparrow 1). \tag{101}$$

Theorem 2 follows then directly by combining (99), (101) and standard algebraic manipulations, recalling that without loss of generality we focused on the waiting-time distributions Q_1 . $\square$