Heavy traffic analysis of a polling model with retrials and glue periods

Abidini, M.A.; Dorsman, J.-P.; Resing, J.

Link to publication

Taylor & Francis
Taylor & Francis Group

Heavy traffic analysis of a polling model with retrials and glue periods

Murtuza Ali Abidini[a], Jan-Pieter Dorsman[b], and Jacques Resing[a]

[a]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands; [b]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

We present a heavy traffic analysis of a single-server polling model, with the special features of retrials and glue periods. The combination of these features in a polling model typically occurs in certain optical networking models, and in models where customers have a reservation period just before their service period. Just before the server arrives at a station there is some deterministic glue period. Customers (both new arrivals and retrials) arriving at the station during this glue period will be served during the visit of the server. Customers arriving in any other period leave immediately and will retry after an exponentially distributed time. As this model defies a closed-form expression for the queue length distributions, our main focus is on their heavy-traffic asymptotics, both at embedded time points (beginnings of glue periods, visit periods, and switch periods) and at arbitrary time points. We obtain closed-form expressions for the limiting scaled joint queue length distribution in heavy traffic. We show that these results can be used to accurately approximate the performance of the system for the complete spectrum of load values by use of interpolation approximations.

1. Introduction

Polling models are queueing models in which a single server, alternatingly, visits a number of queues in some prescribed order. These models have been extensively studied in the literature. For example, various different service disciplines (rules which describe the server's behavior while visiting a queue) and both models with and without switch-over times have been considered. We refer to Takagi[22,23] and Vishnevskii and Semenova[24] for some literature reviews and to Boon, van der Mei and Winands[3], Levy and Sidi[13] and Takagi[21] for overviews of the applicability of polling models.

Motivated by questions regarding the performance modeling and analysis of optical networks, the study of polling models with *retrials* and *glue periods* was initiated in Boxma and Resing[5]. In a communication network, packets must be routed from source to destination, passing through a series of nodes and a protocol decides which packet may be transmitted at these points. A cyclic polling strategy, cyclic meaning that there is a fixed pattern for giving service to particular ports/stations, is used as a protocol here. The "wired" part of communication networks these days is almost completely replaced by optical networks. These networks utilize optical fiber cables as the primary communication medium for transporting data as light pulses (photons) between source and destination. Optical nodes, unlike electronics, have a problem with buffering of optical packets, as photons cannot be stopped. This problem is usually solved by converting a light pulse to electric data, storing it and then reconverting it to a light pulse when the node is ready for further transmission. This is a time and energy consuming process. To overcome this issue, whenever there is a need to buffer photons, they can be forced to move locally in fiber loops. These fiber loops or fiber delay lines (FDL) originate and end at the head of a switch. When a photon arrives at the switch at a time it cannot be served, it is sent into an FDL, thereby incurring a small delay to its time of arrival without getting lost or displaced. Depending on the availability, requirement, traffic, size of photon and other such factors, the length (delay produced) of these FDLs can differ. Hence, we assume that these FDLs delay the photons by a random amount of time. Also, if a packet does not receive service after a cycle through an FDL, then depending on the model it can go into either the same or a longer or a shorter or randomly to any of the available FDLs. Hence, we assume that two consecutive retrials are independent of each other. This FDL feature can be modeled by a retrial queue.

A sophisticated technology that one might try to add to this is varying the speed of light by changing the refractive index of the fiber loop, cf.[16] By increasing the refractive index in a small part of the loop we can achieve "slow light," which implies slowing the packets. Just before a service period at a port starts, the refractive index in a small region at the end of FDLs can be increased, thereby slowing down the packets and "queueing" them, so that they are available for service when the service period starts. This feature is, in our work, modeled as glue periods immediately before the visit period of the corresponding station. Packets (both new arrivals and retrials) arriving in this glue period can be served in that subsequent visit period. Packets arriving in any other period leave immediately and will retry after an exponentially distributed time. These assumptions regarding retrials and glue periods reflect the specific properties of optical buffering, i.e., FDLs and slowing down the packets by varying the refractive index.

Not restricting ourselves to optical networks, one can also interpret a glue period as a reservation period, i.e., a period in which customers can make a reservation at a station for service in the subsequent visit period of that station. The polling models with retrials and reservation periods could be interesting objects of study in, e.g., healthcare. In our model, the reservation period immediately precedes the visit period and could be seen as the last part of a switchover period.

In Ref[5], the joint queue length process is analyzed both at embedded time points (beginnings of glue periods, visit periods and switch-over periods) and at arbitrary time points, for the model with two queues and *deterministic* glue periods. This analysis is later on extended in Abidini, Boxma and Resing[2] to the model with a general number of queues. After that, in Abidini et al.[1], an algorithm is presented to obtain the moments of the number of customers in each station for the model with *exponentially distributed* glue periods. Furthermore, in Ref[1] also a workload decomposition for the model with *generally distributed* glue periods is derived leading to a pseudo-conservation law. The pseudo-conservation law in its turn is used to obtain approximations of the mean waiting times at the different stations. In these articles, however, no analytical expressions for the complete joint distributions have been derived, which is something we aim to do in this article.

In this manuscript, we will study the above-described polling system with retrials and deterministic glue periods in a heavy traffic regime. The reason that we restrict ourselves to deterministic glue periods is that in that case we can use the relation between our polling model and a multitype branching process with immigration as discussed in Ref[2]. This relation enables us to study the heavy-traffic behavior of the process. Optical networks, as a result of the huge bandwidth provided, are not heavily loaded at the core level, but at the access level, the high volatility of traffic can lead to periods at which the system works under heavy load. The behavior of networks in this heavy-loaded period is a motivation for the heavy traffic analysis of a polling model with retrials and glue periods.

More concretely, we will regard the regime where each of the arrival rates is scaled with the same constant, and subsequently the constant approaches from below that value, for which the system is critically loaded. Then, the workload offered to the server is scaled to such a proportion that the queues are on the verge of instability. Many techniques have been used to obtain the heavy traffic behavior of a variety of different polling models. Initial studies of the heavy traffic behavior of polling systems can be found in Coffman, Puhalskii and Reiman[6,7], where the occurrence of a so-called heavy traffic averaging principle is established. This principle implies that, although the total scaled workload in the system tends to a Bessel-type

diffusion in the heavy-traffic regime, it may be considered as a constant during the course of a polling cycle, while the loads of the individual queues fluctuate like a fluid model. It will turn out that this principle will also hold true for this model. Furthermore, in van der Mei[14], several heavy traffic limits have been established by taking limits in known expressions for the Laplace–Stieltjes transform (LST) of the waiting-time distribution. Alternatively, Olsen and van der Mei[17] provide similar results, by studying the behavior of the descendant set approach (a numerical computation method, cf. Konheim, Levy and Srinivasan[10]) in the heavy traffic limit. For the derivation of heavy traffic asymptotics for our model, however, we will use results from branching theory, mainly those presented in Quine[18]. Earlier, these results have resulted in heavy traffic asymptotics for conventional polling models, see van der Mei[15]. We will use the same method as presented in that article, but for a different class of polling systems that models the dynamics of optical networks. In addition, for some steps of the analysis, we will present new and straightforward proofs, while other steps require a different approach. Furthermore, we will derive asymptotics for the *joint* queue length process at arbitrary time points, as opposed to just the marginal processes as derived in Ref[15]. Due to the additional intricacies of the model at hand, we will need to overcome many arising complex difficulties, as will become apparent later.

The rest of the article is organized as follows. In Section 2, we introduce some notation and present a theorem from Ref[18] on multitype branching processes with immigration. In Section 3, we describe in detail the polling model with retrials and glue periods and recall from Ref[2] how the joint queue length process at some embedded time points in this model is related to multitype branching processes with immigration. Next, we will derive heavy traffic results for our model. In Section 4, we consider the joint queue length process at the start of glue periods. In Section 5, we look at the joint queue length process at the start of visit and switch-over periods, while in Section 6, we consider the joint queue length process at arbitrary time points. Finally, in Section 7, we show how the heavy-traffic results, in combination with a light-traffic result, can be used to approximate performance measures for stable systems with arbitrary system loads.

2. Multitype branching processes with immigration

To derive heavy-traffic results for the model under study, we regard its queue length process as a multitype branching process with immigration. To this end, before introducing the actual model in detail, we will state an important general result from Ref[18] on multitype branching processes with immigration in this section, which we will make significant use of in

the sequel of this article. To state this result, we will first need some notation.

A multitype branching process with immigration has two kinds of individuals: immigrants and offspring. The immigrants in the model are represented by the generating function

$$g(\underline{z}) = \sum_{j_1,\ldots,j_N \geq 0} q(j_1,\ldots,j_N) z_1^{j_1} \ldots z_N^{j_N}.$$

Here, $\underline{z} = (z_1, z_2, \ldots, z_N)$ and $|z_i| \leq 1$, for all $i = 1, \ldots, N$, and $q(j_1, \ldots, j_N)$ is the probability that j_k type-k individuals immigrate into the system in a given generation, for all $k = 1, \ldots, N$. We use this to define the mean immigration vector $\underline{g} = (g_1, \ldots, g_N)^T$, where $g_i = \frac{\partial g(\underline{z})}{\partial z_i}\big|_{\underline{z}=\underline{1}}$, for all $i = 1, \ldots, N$, where $\underline{1}$ represents a vector of which each of the entries equals one.

Similarly, the offspring in the model is represented by the vector of generating functions $h(\underline{z}) = (h_1(\underline{z}), h_2(\underline{z}), \ldots, h_N(\underline{z}))$. Here,

$$h_i(\underline{z}) = \sum_{j_1,\ldots,j_N \geq 0} p_i(j_1,\ldots,j_N) z_1^{j_1} \ldots z_N^{j_N},$$

where $p_i(j_1, \ldots, j_N)$ is the probability that a type-i individual produces j_k type-k individuals, for all $i = 1, \ldots, N$ and $k = 1, \ldots, N$. We use this to define the mean matrix $\mathbf{M} = (m_{i,j})$, where $m_{i,j} = \frac{\partial h_i(\underline{z})}{\partial z_j}\big|_{\underline{z}=\underline{1}}$, for all $i, j = 1, \ldots, N$. The elements $m_{i,j}$ represent the mean number of type-j children produced by a type-i individual per generation. We also define the second-order derivative matrix $K^{(i)} = (k_{j,k}^{(i)})$ where $k_{j,k}^{(i)} = \frac{\partial^2 h_i(\underline{z})}{\partial z_j \partial z_k}\big|_{\underline{z}=\underline{1}}$, for all $i, j, k = 1, \ldots, N$.

Define $\underline{w} = (w_1, \ldots, w_N)^T$ as the normalized right eigenvector corresponding to the maximal eigenvalue ξ of $\mathbf{M}$. Then,

$$\mathbf{M}\underline{w} = \xi\underline{w} \qquad \text{and} \qquad \underline{w}^T\underline{1} = 1.$$

Furthermore, we define $\underline{v} = (v_1, \ldots, v_N)^T$ as the left eigenvector of $\mathbf{M}$, corresponding to the maximal eigenvalue ξ, normalized such that

$$\underline{v}^T\underline{w} = 1.$$

Additionally, we give the following general notation in order to state the result of Ref[18]. Any variable x which is dependent on ξ will be denoted by $\hat{x}$ to indicate that it is evaluated at $\xi = 1$. Further, for $0 < \xi < 1$ let

$$\pi_0(\xi) := 0 \qquad \text{and} \qquad \pi_n(\xi) := \sum_{r=1}^{n} \xi^{r-2}, \qquad n = 1, 2, \ldots. \qquad (2.1)$$

We denote with $\Gamma(\alpha, \mu)$ a gamma-distributed random variable. For $\alpha, \mu, x > 0$, its probability density function is given by

$$f(x) = \frac{\mu^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\mu x}, \qquad \text{where} \qquad \Gamma(\alpha) := \int_{t=0}^{\infty} t^{\alpha-1} e^{-t} dt.$$

Now that the required notation is defined, we state the following important result, which we will make use of in the sequel to derive heavy-traffic asymptotics for the polling model with retrials and glue periods. This result is given and proved in Ref[18] (Theorem 4), and it implies that, under certain assumptions on the immigration function $g(\underline{z})$ and the offspring generating function $h(\underline{z})$ (see Eqs. (1.1) and (1.2) in Ref[18]),

$$\frac{1}{\pi(\xi)} \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix} \xrightarrow{d} A \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_N \end{pmatrix} \Gamma(\alpha, 1), \quad \text{when} \quad \xi \uparrow 1. \tag{2.2}$$

Here $\xrightarrow{d}$ means convergence in distribution, $\pi(\xi) := \lim_{n\to\infty} \pi_n(\xi), \alpha := \frac{1}{A} \underline{\hat{g}}^T \underline{\hat{w}}$ and $A := \frac{1}{2} \sum_{i=1}^{N} \hat{v}_i (\underline{\hat{w}}^T \hat{K}^{(i)} \underline{\hat{w}}) > 0$. The vector $(Z_1, Z_2, \ldots, Z_N)$ is defined such that Z_i is the steady-state number of individuals of type-i in the multitype branching process with immigration, for all $i = 1, \ldots, N$.

3. Polling model with retrials and glue periods

In this section, we first define the polling model with retrials and glue periods. Then, we recall from Ref[2] its property that the joint queue length process at the start of glue periods of a certain queue is a multitype branching process with immigration.

3.1. Model description

We consider a single server polling model with multiple queues, Q_i, $i = 1, \ldots, N$. Customers arrive at Q_i according to a Poisson process with rate λ_i; they are called type-i customers. The service times at Q_i are i.i.d., with B_i denoting a generic service time of which the first three[1] moments are finite, with distribution $B_i(\cdot)$ and LST $\tilde{B}_i(\cdot)$. The server cyclically visits all the queues, thus after a visit of Q_i, it switches to $Q_{i+1}, i = 1, \ldots, N$. Successive switch-over times from Q_i to Q_{i+1} are i.i.d., where S_i denotes a generic switch-over time of which the first two moments are finite[1], with distribution $S_i(\cdot)$ and LST $\tilde{S}_i(\cdot)$. We make all the usual independence assumptions about interarrival times, service times and switch-over times at the queues. After a switch of the server to Q_i, there first is a deterministic (i.e., constant) glue period G_i, before the visit of the server at Q_i begins. The significance of the glue period stems from the following assumption.

---

[1]The assumptions of the first three service time moments and the first two switch-over time moments being finite are made for technical purposes. These assumptions are sufficient to satisfy the conditions given in Eqs. (1.1) and (1.2) of Ref[18] which are used in the proof of Theorem 4 of Ref[18], as will become apparent in Lemma 6. As mentioned in Section 2, this theorem plays a key role in our analysis.

Customers who arrive at Q_i do not receive service immediately. When customers arrive at Q_i during a glue period G_i, they stick, joining the queue of Q_i. When they arrive in any other period, they immediately leave and enter into an orbit from which they retry after retrial intervals which are independent of everything else, and exponentially distributed with parameter ν_i, $i = 1, \ldots, N$.

Since customers will only "stick" during the glue period, the service discipline at all queues can be interpreted as being gated. That is, during the visit period at Q_i, the server serves all "glued" customers in that queue, i.e., all type-i customers waiting at the end of the glue period, but none of those in orbit, and neither any new arrivals. We are interested in the steady-state behavior of this polling model with retrials. We hence assume that the stability condition $\rho = \sum_{i=1}^{N} \rho_i < 1$ holds, where $\rho_i := \lambda_i \mathbb{E}[B_i]$.

Note that now the server has three different periods at each station, a deterministic glue period during which customers are glued for service, followed by a visit period during which all the glued customers are served and a switch-over period during which the server moves to the next station. We denote, for $i = 1, \ldots, N$, by $(X_1^{(i)}, X_2^{(i)}, \ldots, X_N^{(i)})$, $(Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_N^{(i)})$ and $(Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_N^{(i)})$ vectors with as distribution the limiting distribution of the number of customers of the different types in the system at the start of a glue period, a visit period and a switch-over period of station i, respectively. Furthermore, we denote, for $i = 1, \ldots, N$, by $(V_1^{(i)}, V_2^{(i)}, \ldots, V_N^{(i)})$ the vector with as distribution the limiting distribution of the number of customers of the different types in the system at an arbitrary point in time during a visit period of station i. During glue and visit periods, we furthermore distinguish between those customers who are queueing in Q_i and those who are in orbit for Q_i. Therefore, we write $Y_i^{(i)} = Y_i^{(iq)} + Y_i^{(io)}$ and $V_i^{(i)} = V_i^{(iq)} + V_i^{(io)}$, for all $i = 1, \ldots, N$, where q represents queueing and o represents in orbit. Finally, we denote by $(L^{(1q)}, \ldots, L^{(Nq)}, L^{(1o)}, \ldots, L^{(No)})$ the vector with as distribution the limiting distribution of the number of customers of the different types in the queue and in the orbit at an arbitrary point in time.

The generating function of the vector of numbers of arrivals at Q_1 to Q_N during a type-i service time B_i is $\beta_i(\underline{z}) := \tilde{B}_i(\sum_{j=1}^{N} \lambda_j(1-z_j))$. Similarly, the generating function of the vector of numbers of arrivals at Q_1 to Q_N during a type-i switch-over time S_i is $\sigma_i(\underline{z}) := \tilde{S}_i(\sum_{j=1}^{N} \lambda_j(1-z_j))$.

3.2. Relation with multitype branching processes

We now identify the relation of the polling model as defined in Section 3.1 with a multitype branching process. In Ref[2], it is shown that the number of customers of different types in the system at the start of a glue period of

station 1 in the polling model with retrials and glue periods is a multitype branching process with immigration. Here, type-i individuals in the branching process represent customers of type-i in orbit in the polling model. The different generations in the branching process correspond to the successive cycles in the polling models. The immigration in a certain generation represents new arrivals during switchover times and glue periods in a certain cycle and/or descendants of these arrivals in the current cycle (corresponding to customers arriving during the service time of these new arrivals) if the new arrivals are served during the current cycle. In particular, it is derived in Ref[2] that the joint probability generating function (PGF) of $X_1^{(1)}, \ldots, X_N^{(1)}$ satisfies

$$\mathbb{E}\left[z_1^{X_1^{(1)}} z_2^{X_2^{(1)}} \ldots z_N^{X_N^{(1)}}\right] = \prod_{i=1}^{N} \sigma^{(i)}(\underline{z}) \prod_{i=1}^{N} e^{-G_i D_i(\underline{z})} \mathbb{E}\left[\left[h_1(\underline{z})\right]^{X_1^{(1)}} \left[h_2(\underline{z})\right]^{X_2^{(1)}} \ldots \left[h_N(\underline{z})\right]^{X_N^{(1)}}\right],$$

(3.1)

where
$$\sigma^{(i)}(\underline{z}) := \sigma_i\big(z_1, \ldots, z_i, h_{i+1}(\underline{z}), \ldots, h_N(\underline{z})\big),$$

$$D_i(\underline{z}) := \sum_{j=1}^{i-1} \lambda_j(1-z_j) + \lambda_i\left(1 - \beta^{(i)}(\underline{z})\right) + \sum_{j=i+1}^{N} \lambda_j\big(1-h_j(\underline{z})\big),$$

$$\beta^{(i)}(\underline{z}) := \beta_i\big(z_1, \ldots, z_i, h_{i+1}(\underline{z}), \ldots, h_N(\underline{z})\big),$$

$$h_i(\underline{z}) := f_i\big(z_1, \ldots, z_i, h_{i+1}(\underline{z}), \ldots, h_N(\underline{z})\big),$$

and
$$f_i(\underline{z}) := \left(1-e^{-v_i G_i}\right)\beta_i(\underline{z}) + e^{-v_i G_i} z_i.$$

The first two factors in (3.1) represent the immigration part of the process. Therefore, we have the immigrant generating function given by

$$g(\underline{z}) = \prod_{i=1}^{N} \sigma^{(i)}(\underline{z}) \prod_{i=1}^{N} e^{-G_i D_i(\underline{z})}.$$

The third factor represents the branching part of the process. Recall that the vector of offspring generating functions is given by

$$h(\underline{z}) = \big(h_1(\underline{z}), h_2(\underline{z}), \ldots, h_N(\underline{z})\big).$$

A customer of type-i present at the start of a glue period of station 1 is effectively replaced by a population with joint PGF $h_i(z)$ in the next cycle.

As explained in detail in Ref[2], (3.1) consists of the product of three factors:

- $\prod_{i=1}^{N} \sigma^{(i)}(\underline{z})$ represents new arrivals during switch-over times and descendants of these arrivals in the current cycle.
- $\prod_{i=1}^{N} e^{-G_i D_i(\underline{z})}$ represents new arrivals during glue periods and descendants of these arrivals in the current cycle. The function $D_i(\underline{z})$ is itself a sum of three terms:

- $\sum_{j=1}^{i-1}\lambda_j(1-z_j)$ represents the arrivals of type $j < i$; these arrivals are not served in the current cycle.
- $\lambda_i(1 - \beta^{(i)}(\underline{z}))$ represents descendants of the arrivals of type-i; these arrivals are all served during the visit of station i in the current cycle.
- $\sum_{j=i+1}^{N}\lambda_j(1-h_j(\underline{z}))$ represents the arrivals or descendants of arrivals of type $j > i$; these arrivals are either served (with probability $1-e^{-v_iG_i}$) or not served (with probability $e^{-v_iG_i}$) in the current cycle.
- $\mathbb{E}[[h_1(\underline{z})]^{X_1^{(1)}}[h_2(\underline{z})]^{X_2^{(1)}}\ldots[h_N(\underline{z})]^{X_N^{(1)}}]$ represents   descendants   of $(X_1^{(1)},\ldots,X_N^{(1)})$ generated in the current cycle.

We now proceed to further identify the branching process by finding its mean matrix M and the mean immigration vector $\underline{g}$.

3.2.1. Mean matrix of branching process

The elements $m_{i,j}$ of the mean matrix $\mathbf{M}$ of the branching process are given by

$$m_{i,j} = f_{i,j} \cdot 1[j \le i] + \sum_{k=i+1}^{N} f_{i,k}m_{k,j}, \tag{3.2}$$

where $f_{i,j} = \frac{\partial f_i(\underline{z})}{\partial z_j}\Big|_{\underline{z}=\underline{1}}$, and hence

$$f_{i,j} = \begin{cases} (1-e^{-v_iG_i})\lambda_j\mathbb{E}[B_i], & i \ne j, \\ (1-e^{-v_iG_i})\rho_i + e^{-v_iG_i}, & i = j. \end{cases} \tag{3.3}$$

In the heavy traffic analysis of this model, the following lemma will be useful.

Lemma 1.

$$\mathbf{M} = \mathbf{M_1}\cdots\mathbf{M_N}, \tag{3.4}$$

where, for $i = 1, 2, \ldots, N$, we have

$$\mathbf{M_i} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & \vdots & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ f_{i,1} & f_{i,2} & \cdots & f_{i,i-1} & f_{i,i} & f_{i,i+1} & \vdots & \vdots & f_{i,N} \\ 0 & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & 1 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{3.5}$$

Proof. See Appendix.   □

Remark 1. (Intuition behind Lemma 1) The matrix $\mathbf{M_i}$ represents what happens with customers during a visit period at station i. Customers at station i itself are either served or not served, leading to the *ith* row with elements $f_{i,j}$. Customers at all other stations are not served leading to 1's on the diagonal and 0's outside the diagonal. We obtain the product $\mathbf{M_1} \cdots \mathbf{M_N}$ because a cycle consists successively of visit periods of station 1, station 2, ..., up to station N.

3.2.2. Mean number of immigrants

Next, we look at the immigration part of the process. Let g_i be the mean number of type-i individuals which immigrate into the system in each generation. Equation (3.12) of Ref[2] gives us

$$
g_i = \sum_{k=1}^{N} \lambda_k \left(\left(\sum_{j=1}^{k-1} \left(G_j + \mathbb{E}[S_j] \right) \right) \left(1 - e^{-\nu_k G_k} \right) + G_k \right) m_{k,i}
$$
$$
+ \lambda_i \left(\sum_{j=1}^{i-1} \left(G_j + \mathbb{E}[S_j] \right) e^{-\nu_i G_i} + \sum_{j=i}^{N} \mathbb{E}[S_j] + \sum_{j=i+1}^{N} G_j \right). \tag{3.6}
$$

The right-hand side of (3.6) is the sum of two terms. The term $\sum_{k=1}^{N} \lambda_k \left(\left(\sum_{j=1}^{k-1} (G_j + \mathbb{E}[S_j]) \right) (1 - e^{-\nu_k G_k}) + G_k \right) m_{k,i}$ represents the mean number of type-i customers which are descendants of customers of type k, arriving during glue periods and switch-over periods before the visit period of station k and served during the visit at station k, in the current cycle. The first part of the second term $\lambda_i \sum_{j=1}^{i-1} (G_j + \mathbb{E}[S_j]) e^{-\nu_i G_i}$ represents the mean number of customers of type i which arrive during glue periods and switch-over periods before the visit of the server at station i and which are not served during the visit of station i in the current cycle. The second part of the second term $\lambda_i (\sum_{j=i}^{N} \mathbb{E}[S_j] + \sum_{j=i+1}^{N} G_j)$ represents the mean number of customers of type i which arrive during glue periods and switch-over periods after the visit period of station i in the current cycle. Note that each of the terms mentioned above is non-negative and finite. Furthermore, for non-zero glue periods and arrival rates, at least one of the terms is non-zero. Therefore, we have $0 < g_i < \infty$.

Remark 2. Note that the branching part of the process only represents descendants of customers which are present in the system at the start of a glue period of station 1. Customers which arrive at stations during glue periods and switch-over periods are not represented by the branching part of the process. Instead they and their descendants are represented by the immigration part of the process. Both glue periods and switch-over periods

can be considered as parts of the cycle during which the server is not working. This rather unexpected feature explains why the polling model at hand is not part of the class of polling models considered by Ref[15], but requires an analysis on its own.

4. Heavy traffic analysis: number of customers at start of glue periods of station 1

Now that we have successfully modeled the polling system as a multitype branching process with immigration, we derive the limiting scaled joint queue length distribution in each station at the start of glue periods of station 1 by following the same line of proof as that of Ref[15]. In Ref[15], the author first proves a couple of lemmas for a conventional branching-type polling system without retrials and glue periods and, afterwards, uses these lemmas to give the heavy-traffic asymptotics of the joint queue length process at certain embedded time points. In the following subsection, we will derive similar lemmas in order to derive a heavy-traffic theorem for our polling system with retrials and glue periods.

Note that when we scale our system such that $\rho \uparrow 1$, we are effectively changing the arrival rate at each station while keeping the service times and the ratios of the arrival rates fixed. Let any variable x which is dependent on ρ be denoted by $\hat{x}$ whenever it is evaluated at $\rho = 1$. Therefore, we have for any system that, $\lambda_i = \rho \hat{\lambda}_i$.

From Theorem 1 of Ref[26], we know that if all elements of a matrix are continuous in some variable, then the real eigenvalues of this matrix are also continuous in that variable. As each element of $\mathbf{M}$ is a continuous function of ρ, the maximal eigenvalue ξ is a continuous function of ρ as well. Furthermore, from Lemmas 3, 4 and 5 of Ref[20] we know that $\xi < 1$ when $\rho < 1$, $\xi = 1$ when $\rho = 1$ and $\xi > 1$ when $\rho > 1$. Therefore, we have that ξ is a continuous function of ρ and

$$\lim_{\rho \uparrow 1} \xi(\rho) = \xi(1) = 1.$$

4.1. Preliminary results and lemmas

Lemma 2. *The normalized right and left eigenvectors of $\hat{\mathbf{M}}$, the mean matrix of the system with $\rho = 1$, corresponding to the maximal eigenvalue $\xi = 1$, are respectively given by*

$$\underline{\hat{w}} = \begin{pmatrix} \hat{w}_1 \\ \vdots \\ \hat{w}_N \end{pmatrix} = \frac{\underline{b}}{|\underline{b}|} \quad and \quad \underline{\hat{v}} = \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_N \end{pmatrix} = \frac{|\underline{b}|}{\delta} \underline{\hat{u}},$$

where

$$
\underline{b} = \begin{pmatrix} \mathbb{E}[B_1] \\ \vdots \\ \mathbb{E}[B_N] \end{pmatrix}, \quad \underline{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}, \quad |\underline{b}| := \sum_{j=1}^{N} \mathbb{E}[B_j],
$$

$$
u_j := \lambda_j \left[\frac{e^{-v_j G_j}}{1 - e^{-v_j G_j}} + \sum_{k=j}^{N} \rho_k \right] \quad and \quad \delta := \underline{\hat{u}}^T \underline{b}.
$$

Proof. See [Appendix](#). □

Remark 3. Alternatively, we could have used Lemma 4 from Ref[15] to find the left and normalized right eigenvectors. The normalized right eigenvector $\hat{w}$ is the same as given in Ref[15]. To find the left eigenvector $\hat{v}$ from Ref[15], we first need to calculate the exhaustiveness factor f_j. In our model, this exhaustiveness factor is given by $f_j = (1-e^{-v_j G_j})(1-\rho_j)$. Each customer of type j, present at the start of a glue period at station j, is served with probability $(1-e^{-v_j G_j})$ and during that service time on average ρ_j new type-j customers will arrive. Furthermore, with probability $e^{-v_j G_j}$ a customer of type j, present at the start of a glue period at station j, is not served. Therefore, we have $1-f_j = (1-e^{-v_j G_j})\rho_j + e^{-v_j G_j}$, and hence the exhaustiveness factor is given by $f_j = (1-e^{-v_j G_j})(1-\rho_j)$. Substituting this exhaustiveness factor in Lemma 4 of Ref[15] we get

$$
\begin{aligned}
u_j &= \lambda_j \left[\frac{(1-\rho_j)\left(1-(1-e^{-v_j G_i})(1-\rho_j)\right)}{(1-e^{-v_j G_j})(1-\rho_j)} + \sum_{k=j+1}^{N} \rho_k \right] \\
&= \lambda_j \left[\frac{e^{-v_j G_j} + (1-e^{-v_j G_j})\rho_j}{1-e^{-v_j G_j}} + \sum_{k=j+1}^{N} \rho_k \right] \\
&= \lambda_j \left[\frac{e^{-v_j G_j}}{1-e^{-v_j G_j}} + \sum_{k=j}^{N} \rho_k \right],
\end{aligned}
$$

which is in agreement with [Lemma 2](#).

Remark 4. In [Lemma 2](#), we have given the left and normalized right eigenvectors for the mean matrix $\hat{M}$ at eigenvalue $\xi = 1$. Note that this mean matrix is defined for the branching process when we consider the beginning of a glue period of station 1 as the initial point of the cycle. Instead, if we consider the beginning of a glue period of station i as the initial point of the cycle, we get, for eigenvalue $\xi = 1$, the same normalized right eigenvector $\hat{w}$. However, the left eigenvector is now given by the vector $\hat{\underline{v}}^{(i)}$

defined by

$$\hat{\underline{v}}^{(i)} = \begin{pmatrix} \hat{v}_1^{(i)} \\ \vdots \\ \hat{v}_N^{(i)} \end{pmatrix} = \frac{|\underline{b}|}{\delta} \hat{\underline{u}}^{(i)},$$

where

$$\underline{u}^{(i)} := \begin{pmatrix} u_1^{(i)} \\ \vdots \\ u_N^{(i)} \end{pmatrix} \quad \text{and} \quad u_j^{(i)} := \begin{cases} \lambda_j \left[\dfrac{e^{-v_j G_j}}{1 - e^{-v_j G_j}} + \sum_{k=j}^{N} \rho_k + \sum_{k=1}^{i-1} \rho_k \right], & i \le j, \\[4mm] \lambda_j \left[\dfrac{e^{-v_j G_j}}{1 - e^{-v_j G_j}} + \sum_{k=j}^{i-1} \rho_k \right], & i > j. \end{cases}$$

Note that $\delta = \hat{\underline{u}}^{(1)^T} \underline{b} = \hat{\underline{u}}^{(i)^T} \underline{b}$ and $u_j^{(1)} = u_j^{(N+1)} = u_j$, for all $i, j = 1, \dots, N$. In this article, we prove all the lemmas and theorems using $\hat{v} = \hat{v}^{(1)}$. However, we can instead use $\hat{v}^{(i)}$ and prove the same by just changing the initial point of the cycle from the beginning of a glue period of station 1 to the beginning of a glue period of station i.

In Lemma 2, we have evaluated the normalized right, and left eigenvectors of $\mathbf{M}$ at the maximal eigenvalue, when $\rho \uparrow 1$. We will now use this to compute the value of the derivative of this eigenvalue as $\rho \uparrow 1$.

Lemma 3. *For the maximal eigenvalue $\xi = \xi(\rho)$ of the matrix $\mathbf{M}$, the derivative of $\xi(\rho)$ w.r.t. ρ satisfies*

$$\xi'(1) = \frac{1}{\delta}.$$

Proof. See Appendix. □

For the result in (2.2), we need all the second-order derivatives $\frac{\delta^2 h_i(\underline{z})}{\delta z_j \delta z_k}$ of the function $h_i(\underline{z})$. In Lemma 4, we first find $\frac{\delta^2 h_i(\underline{z})}{\delta z_j \delta z_k}$, for all i, j and k, and then use them to find the parameter A as defined in (2.2).

Lemma 4. *For the second-order derivative matrix $K^{(i)} = (k_{j,k}^{(i)})$ where $k_{j,k}^{(i)} = \frac{\delta^2 h_i(\underline{z})}{\delta z_j \delta z_k}\Big|_{\underline{z}=\underline{1}}$, for all $i, j, k = 1, \dots, N$, we have that*

$$A := \frac{1}{2} \sum_{i=1}^{N} \hat{v}_i^{(1)} \left(\hat{\underline{w}}^T \hat{K}^{(i)} \hat{\underline{w}} \right) = \frac{1}{2\delta |\underline{b}|} \frac{b^{(2)}}{b^{(1)}},$$

where

$$b^{(j)} = \frac{\sum_{i=1}^{N} \lambda_i \mathbb{E}\left[B_i^j\right]}{\sum_{i=1}^{N} \lambda_i},$$

for j = 1, 2.

Proof. See Appendix. □

At this point, we have determined all parameters required to deploy in (2.2), except for the constant α. This parameter depends on the immigration part of our process and is given by the following lemma.

Lemma 5. *For $\underline{g} = (g_1, \ldots, g_N)^T$, we have that*

$$\alpha := \frac{1}{A} \hat{\underline{g}}^T \hat{\underline{w}} = 2r\delta \frac{b^{(1)}}{b^{(2)}}, \tag{4.1}$$

where

$$r = \sum_{i=1}^{N} \left(\mathbb{E}[S_i] + G_i\right).$$

Proof. See Appendix. □

Now that we have determined all the parameters of (2.2), we give a final lemma in which we show that the multitype branching process defined by our polling system with retrials and glue periods actually falls in the framework put forward in Ref[18]. In particular, we show that the offspring generating function $h(\underline{z})$ falls in the class K as defined in (1.1) of Ref[18] and that the immigration generating function $g(\underline{z})$ falls in the class J as defined by (1.2) of Ref[18].

Lemma 6. *The generating functions for offspring and immigration, $h(\underline{z})$ and $g(\underline{z})$, respectively satisfy the conditions defined in (1.1) and (1.2) of [18].*

Proof. See Appendix. □

4.2. The heavy traffic theorem

Similar to the procedure used in Ref[15], we will now combine the preliminary work in Section 4.1 with Theorem 4 in [18] in order to obtain the following heavy traffic theorem for the complete queue length process at cycle starts.

Theorem 1. *For the cyclic polling system with retrials and glue periods, the scaled steady-state joint queue length vector at the start of glue periods at station 1 satisfies*

$$(1-\rho)\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \xrightarrow[d]{\rho\uparrow 1} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\begin{pmatrix} \hat{u}_1^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix}\Gamma(\alpha,1), \qquad (4.2)$$

where

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

Proof. In Lemma 6, we have shown that the branching process underlying the polling model with retrials and glue periods fits the framework of Ref[18]. As a result, it now follows from (2.2) that

$$\frac{1}{\pi\big(\xi(\rho)\big)}\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \xrightarrow[d]{} A\begin{pmatrix} \hat{v}_1^{(1)} \\ \vdots \\ \hat{v}_N^{(1)} \end{pmatrix}\Gamma(\alpha,1), \quad \text{when} \quad \rho\uparrow 1, \qquad (4.3)$$

where $\pi(\xi(\rho)) := \lim_{n\to\infty}\pi_n(\xi(\rho))$, and A and $\underline{\hat{v}}^{(1)}$ and $\alpha = \frac{1}{A}\underline{\hat{g}}^T\underline{\hat{w}}$, are as defined in Lemmas 2, 4 and 5.

From (2.1) we can say that, for $\rho < 1$,

$$\pi\big(\xi(\rho)\big) = \frac{1}{\xi(\rho)\big(1 - \xi(\rho)\big)}.$$

Using this, together with Lemma 3, gives

$$\lim_{\rho\uparrow 1}(1-\rho)\pi\big(\xi(\rho)\big) = \lim_{\rho\uparrow 1}\frac{1-\rho}{\xi(\rho)\big(1-\xi(\rho)\big)} = \lim_{\rho\uparrow 1}\frac{-1}{\xi'(\rho)\big(1-2\xi(\rho)\big)} = \lim_{\rho\uparrow 1}\frac{1}{\xi'(\rho)} = \delta.$$
$$(4.4)$$

Therefore, multiplying and dividing the LHS of (4.3) with $1-\rho$, we get

$$\frac{1-\rho}{(1-\rho)\pi\big(\xi(\rho)\big)}\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \xrightarrow[d]{} A\begin{pmatrix} \hat{v}_1^{(1)} \\ \vdots \\ \hat{v}_N^{(1)} \end{pmatrix}\Gamma(\alpha,1), \quad \text{when} \quad \rho\uparrow 1.$$

Using (4.4), this gives

$$(1-\rho)\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \xrightarrow[d]{} \frac{1}{2|\underline{b}|}\frac{b^{(2)}}{b^{(1)}}\begin{pmatrix} \hat{v}_1^{(1)} \\ \vdots \\ \hat{v}_N^{(1)} \end{pmatrix}\Gamma(\alpha,1), \quad \text{when} \quad \rho\uparrow 1,$$

and hence

$$(1-\rho)\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_N^{(1)} \end{pmatrix} \underset{d}{\to} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\begin{pmatrix} \hat{u}_1^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix}\Gamma(\alpha,1), \quad \text{when} \quad \rho \uparrow 1.$$

4.3. Discussion of results: connection with a binomially gated polling model

It turns out that the heavy-traffic results that we obtained in this section for the model at hand, are similar to those of a binomially gated polling model (see e.g., Ref[12]). The dynamics of the binomially gated polling model are much like those of a conventional gated polling model, except that after dropping a gate at Q_i, the customers before it will each be served in the corresponding visit period with probability p_i in an i.i.d. way, rather than with probability one as in the gated model. In particular, the heavy traffic analysis of our model coincides with that of a binomially gated polling model with the same interarrival time distributions, service time distributions and switch-over time distributions, and probability parameters $p_i = 1-e^{-v_i G_i}$. To check this, we note that the binomially gated polling model with these probability parameters falls within the framework of the seminal work of Ref[15] when taking the exhaustiveness parameters $f_i = (1-\rho_i)(1-e^{-v_i G_i})$, after which it is easily verified that Theorem 5 of Ref[15] coincides with Theorem 1. Note, however, that although we also exploit a branching framework in this paper, the model considered in this article does not fall directly in the class of polling models considered in Ref[15], due to the intricate immigration dynamics it exposes.

The intuition behind this remarkable connection is as follows. First, we have that a binomially gated polling model does not have the feature of glue periods. However, in a heavy-traffic regime, the server in our model will reside in a visit period for 100% of the time, so that glue periods hardly occur in this regime either. Furthermore, in a binomially gated polling model, each customer present at the start of a visit period at Q_i will be served within that visit period with probability $p_i = 1-e^{-v_i G_i}$ in an i.i.d. fashion. Note that something similar happens with the model at hand. There, the start of a visit period coincides with the conclusion of a glue period. During this glue period, all customers present in the orbit of the queue will, independently from one another, queue up for the next visit period with probability $1-e^{-v_i G_i}$. These two facts explain the analogy.

It is worth to emphasize that this analogy, remarkable though it is, does not help us in the further analysis towards the asymptotics of the customer population at an arbitrary point in time. While Theorem 1 is now aligned with Theorem 5 of Ref[15], we cannot use the subsequent analysis steps in

that article to get to results concerning the customer population in heavy traffic at an arbitrary point in time. This is much due to the fact that the strategy of Ref[15] exploits a relation between the queue length of Q_1 at a cycle start and the virtual waiting time of that queue at an arbitrary point in time. Since the type-i customers in our model are not served in the order of arrival, as is usually assumed, such a relation is hard to derive and is essentially unknown. As an alternative, we will extend the current heavy traffic asymptotics at cycle starts to certain other embedded epochs in Section 5, and eventually to arbitrary points in time in Section 6.

5. Heavy traffic analysis: number of customers at other embedded time points

A cycle in the polling system with retrials and glue periods passes through three different phases: glue periods, visit periods and switch-over periods. In the previous section, in Theorem 1, we studied the behavior of the scaled steady-state joint queue length vector at the start of glue periods at station 1. We will now extend this result to the scaled steady-state joint queue length vector at the start of a visit period and the start of a switch-over period in Theorems 2 and 3.

Theorem 2. *For the cyclic polling system with retrials and glue periods, the scaled steady-state joint queue length vector at the start of visit periods at station 1 satisfies*

$$(1-\rho)\begin{pmatrix} Y_1^{(1q)} \\ Y_1^{(1o)} \\ Y_2^{(1)} \\ \vdots \\ Y_N^{(1)} \end{pmatrix} \xrightarrow{d} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\begin{pmatrix} (1-e^{-\nu_1 G_1})\hat{u}_1^{(1)} \\ e^{-\nu_1 G_1}\hat{u}_1^{(1)} \\ \hat{u}_2^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix}\Gamma(\alpha,1), \quad \text{when} \quad \rho \uparrow 1.$$

$$(5.1)$$

Proof. The distribution of the number of new customers of type j entering the system during a glue period of station i is stochastically smaller than that of the number of events $G_j^{(i)}$ in a Poisson process with rate $\hat{\lambda}_j$ during an interval of length G_i. This is due to the fact that the arrival rate $\lambda_j = \rho\hat{\lambda}_j$ does not exceed $\hat{\lambda}_j$. Since $G_j^{(i)}$ is finite with probability 1, we have that $(1-\rho)G_j^{(i)} \to 0$ with probability 1, as $\rho \uparrow 1$. Therefore the limiting scaled joint queue length distribution, for all customers other than type i, at the start of a glue period is the same as at the start of a visit period of station i.

Furthermore, the $X_i^{(i)}$ customers of type i, present in the system at the start of a glue period of station i, join the queue, independently of each

other, with probability $1-e^{-v_iG_i}$ during the glue period. Let $\{U_i, i \geq 0\}$ be a series of i.i.d. random variables where U_k indicates whether the k-th customer joins the queue or stays in orbit, for all $k = 1, \ldots, X_i^{(i)}$. More specifically, $U_k = 1$ if the customer joins the queue, with probability $1-e^{-v_iG_i}$, and $U_k = 0$ if the customer stays in orbit, with probability $e^{-v_iG_i}$. Then the number of customers of type i in the queue ($Y_i^{(iq)}$) and in the orbit ($Y_i^{(io)}$) at the start of a visit period at station i are given by

$$Y_i^{(iq)} = \sum_{k=1}^{X_i^{(i)}} U_k \qquad \text{and} \qquad Y_i^{(io)} = X_i^{(i)} - \sum_{k=1}^{X_i^{(i)}} U_k.$$

Since $X_i^{(i)} \to \infty$ with probability 1, as $\rho \uparrow 1$, we have by virtue of the weak law of large numbers that

$$\frac{Y_i^{(iq)}}{X_i^{(i)}} = \frac{\sum_{k=1}^{X_i^{(i)}} U_k}{X_i^{(i)}} \xrightarrow[\mathbb{P}]{} 1-e^{-v_iG_i}, \qquad \text{when} \quad \rho \uparrow 1, \tag{5.2}$$

where $\xrightarrow{\mathbb{P}}$ means convergence in probability. Similarly we have

$$\frac{Y_i^{(io)}}{X_i^{(i)}} = \frac{X_i^{(i)} - \sum_{k=1}^{X_i^{(i)}} U_k}{X_i^{(i)}} \xrightarrow[\mathbb{P}]{} e^{-v_iG_i}, \qquad \text{when} \quad \rho \uparrow 1. \tag{5.3}$$

Therefore, using Slutsky's convergence theorem[9], along with (4.3), (5.2) and (5.3) and the arguments above, we get

$$(1-\rho)\begin{pmatrix} Y_1^{(1q)} \\ Y_1^{(1o)} \\ Y_2^{(1)} \\ \vdots \\ Y_N^{(1)} \end{pmatrix} \xrightarrow[d]{\rho \uparrow 1} \frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \begin{pmatrix} (1-e^{-v_1G_1})\hat{u}_1^{(1)} \\ e^{-v_1G_1}\hat{u}_1^{(1)} \\ \hat{u}_2^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix} \Gamma(\alpha, 1), \qquad \text{when} \quad \rho \uparrow 1.$$

We end this section by considering the scaled steady-state joint queue length vector at the start of a switch-over period from station 1 to station 2.

Theorem 3. *For the cyclic polling system with retrials and glue periods, the scaled steady-state joint queue length vector at the start of a switch-over period from station 1 to station 2 satisfies*

$$(1-\rho)\begin{pmatrix} Z_1^{(1)} \\ Z_2^{(1)} \\ \vdots \\ Z_N^{(1)} \end{pmatrix} \xrightarrow[d]{} \frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \begin{pmatrix} e^{-v_1 G_1}\hat{u}_1^{(1)} & +(1-e^{-v_1 G_1})\hat{u}_1^{(1)}\hat{\lambda}_1 \mathbb{E}[B_1] \\ \hat{u}_2^{(1)} & +(1-e^{-v_1 G_1})\hat{u}_1^{(1)}\hat{\lambda}_2 \mathbb{E}[B_1] \\ \vdots \\ \hat{u}_N^{(1)} & +(1-e^{-v_1 G_1})\hat{u}_1^{(1)}\hat{\lambda}_N \mathbb{E}[B_1] \end{pmatrix} \Gamma(\alpha,1), \quad \text{when } \rho \uparrow 1.$$

(5.4)

Proof. The number of customers in the orbit of station j at the start of a switch-over period from station i to station $i+1$ equals the number of customers in the orbit at the start of the visit of station i plus the Poisson arrivals with rate λ_j, during the service of customers in the queue of station i, say $J_j^{(i)}$. In other words, we have that

$$Z_j^{(i)} = \begin{cases} Y_j^{(i)} + J_j^{(i)}, & j \neq i, \\ Y_i^{(io)} + J_i^{(i)}, & j = i. \end{cases} \tag{5.5}$$

Note that $J_j^{(i)}$ is the sum of Poisson arrivals with rate λ_j during $Y_i^{(iq)}$ independent service times with distribution B_i. Let $D_{i,j,k}$ be the number of Poisson arrivals with rate λ_j during the kth service in the visit period of station i. Thus

$$J_j^{(i)} = \sum_{k=1}^{Y_i^{(iq)}} D_{i,j,k}.$$

Since $Y_i^{(iq)} \to \infty$ as $\rho \uparrow 1$, and $\mathbb{E}[B_i]$ is finite, we have by virtue of the weak law of large numbers that

$$\frac{J_j^{(i)}}{Y_i^{(iq)}} \xrightarrow[\mathbb{P}]{} \hat{\lambda}_j \mathbb{E}[B_i], \quad \text{when } \rho \uparrow 1. \tag{5.6}$$

Therefore, using Slutsky's convergence theorem along with (5.1), (5.5) and (5.6) we get

$$(1-\rho)\begin{pmatrix} Z_1^{(1)} \\ Z_2^{(1)} \\ \vdots \\ Z_N^{(1)} \end{pmatrix} \xrightarrow[d]{} \frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \begin{pmatrix} e^{-v_1 G_1}\hat{u}_1^{(1)} & +(1-e^{-v_1 G_1})\hat{u}_1^{(1)}\hat{\lambda}_1 \mathbb{E}[B_1] \\ \hat{u}_2^{(1)} & +(1-e^{-v_1 G_1})\hat{u}_1^{(1)}\hat{\lambda}_2 \mathbb{E}[B_1] \\ \vdots \\ \hat{u}_N^{(1)} & +(1-e^{-v_1 G_1})\hat{u}_1^{(1)}\hat{\lambda}_N \mathbb{E}[B_1] \end{pmatrix} \Gamma(\alpha,1), \quad \text{when } \rho \uparrow 1.$$

Remark 5. Alternatively, Theorems 2 and 3 can be obtained by exploiting known relations between the joint PGFs of the vectors $(X_1^{(1)}, \ldots, X_N^{(1)})$, $(Y_1^{(1q)}, Y_1^{(1o)}, Y_2^{(1)} \ldots, Y_N^{(1)})$ and $(Z_1^{(1)}, \ldots, Z_N^{(1)})$ given in Equations (3.2) and (3.3) of Ref[2]. After replacing each parameter z_j in these functions by $z_j^{(1-\rho)}$ and taking the limit of ρ going to one from below, these expressions give the relations between the joint PGFs of the heavy traffic

distributions. Combining these results with Theorem 1 and subsequently invoking Levy's continuity theorem (see e.g., Section 18.1 of [25]) then readily imply the theorems.

Remark 6. Throughout this section, we have focused on the joint queue length process at the start of a glue, visit or switch-over period at Q_1. However, similar results for the starts of these periods at any Q_i can be obtained by either simply reordering indices, or by exploiting the relations obtained in Ref[2] between $(X_1^{(i)}, \ldots, X_N^{(i)}), (Y_1^{(iq)}, Y_1^{(io)}, Y_2^{(i)} \ldots, Y_N^{(i)}), (Z_1^{(i)}, \ldots, Z_N^{(i)})$ and $(X_1^{(i+1)}, \ldots, X_N^{(i+1)})$.

6. Heavy traffic analysis: number of customers at arbitrary time points

In this section, we look at the limiting scaled joint queue length distribution of the number of customers at the different stations at an arbitrary time point. At such a point in time, the system can be in the glue period, the visit period or the switch-over period of some station i, with probability $\frac{(1-\rho)G_i}{\sum_{i=1}^N (G_i+\mathbb{E}[S_i])}$, ρ_i and $\frac{(1-\rho)\mathbb{E}[S_i]}{\sum_{i=1}^N (G_i+\mathbb{E}[S_i])}$ respectively. As $\rho \uparrow 1$, the probabilities $\frac{(1-\rho)G_i}{\sum_{i=1}^N (G_i+\mathbb{E}[S_i])}$ and $\frac{(1-\rho)\mathbb{E}[S_i]}{\sum_{i=1}^N (G_i+\mathbb{E}[S_i])}$ both converge to 0. Therefore, we only need to study the scaled steady-state joint queue length vector at an arbitrary time in each of the N visit periods.

Theorem 4. *For the cyclic polling system with retrials and glue periods, the scaled steady-state joint queue length vector at an arbitrary time point in a visit period of station 1 satisfies*

$$(1-\rho)\begin{pmatrix} V_1^{(1q)} \\ V_1^{(1o)} \\ V_2^{(1)} \\ \vdots \\ V_N^{(1)} \end{pmatrix} \underset{d}{\to} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\left[\begin{pmatrix} (1-e^{-v_1 G_1})\hat{u}_1^{(1)} \\ e^{-v_1 G_1}\hat{u}_1^{(1)} \\ \hat{u}_2^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix}\right.$$

$$\left. +(1 - e^{-v_1 G_1})\hat{u}_1^{(1)}U\begin{pmatrix} -1 \\ \hat{\lambda}_1 E[B_1] \\ \hat{\lambda}_2 E[B_1] \\ \vdots \\ \hat{\lambda}_N E[B_1] \end{pmatrix}\right]\Gamma(\alpha + 1, 1), \quad \text{when } \rho \uparrow 1. \qquad (6.1)$$

Proof. We will use Equation (3.19) of[2] to prove this. This equation states that the joint generating function, $R_{vi}^{(i)}(\underline{z}_q, \underline{z}_o)$, of the numbers of customers

in the queue and in the orbits at an arbitrary time point in a visit period of Q_i is given by

$$R_{vi}^{(i)}(\underline{z}_q, \underline{z}_o) = \frac{z_{iq}\left(\mathbb{E}\left[z_{iq}^{Y_i^{(iq)}}\left(\prod_{j=1,j\neq i}^N z_{jo}^{Y_j^{(i)}}\right)z_{io}^{Y_i^{(io)}}\right] - \mathbb{E}\left[\tilde{B}_i\left(\sum_{j=1}^N \lambda_j(1-z_{jo})\right)^{Y_i^{(iq)}}\left(\prod_{j=1,j\neq i}^N z_{jo}^{Y_j^{(i)}}\right)z_{io}^{Y_i^{(io)}}\right]\right)}{\mathbb{E}\left[Y_i^{(iq)}\right]\left(z_{iq} - \tilde{B}_i\left(\sum_{j=1}^N \lambda_j(1-z_{jo})\right)\right)}$$

$$\times \frac{1-\tilde{B}_i\left(\sum_{j=1}^N \lambda_j(1-z_{jo})\right)}{\left(\sum_{j=1}^N \lambda_j(1-z_{jo})\right)\mathbb{E}[B_i]}.$$

Evaluating the above generating function in the points $\tilde{\underline{z}}_q = (z_{1q}^{(1-\rho)}, \ldots, z_{Nq}^{(1-\rho)})$ and $\tilde{\underline{z}}_o = (z_{1o}^{(1-\rho)}, \ldots, z_{No}^{(1-\rho)})$, we get

$$R_{vi}^{(i)}(\tilde{\underline{z}}_q, \tilde{\underline{z}}_o) =$$

$$\frac{z_{iq}^{(1-\rho)}\left(\mathbb{E}\left[z_{iq}^{(1-\rho)Y_i^{(iq)}}\left(\prod_{j=1,j\neq i}^N z_{jo}^{(1-\rho)Y_j^{(i)}}\right)z_{io}^{(1-\rho)Y_i^{(io)}}\right] - \mathbb{E}\left[\tilde{B}_i\left(\sum_{j=1}^N \lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)^{Y_i^{(iq)}}\left(\prod_{j=1,j\neq i}^N z_{jo}^{(1-\rho)Y_j^{(i)}}\right)z_{io}^{(1-\rho)Y_i^{(io)}}\right]\right)}{\mathbb{E}\left[Y_i^{(iq)}\right]\left(z_{iq}^{(1-\rho)} - \tilde{B}_i\left(\sum_{j=1}^N \lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)\right)}$$

$$\times \frac{1-\tilde{B}_i\left(\sum_{j=1}^N \lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)}{\left(\sum_{j=1}^N \lambda_j\left(1 - z_{jo}^{(1-\rho)}\right)\right)\mathbb{E}[B_i]}. \tag{6.2}$$

This equation has two terms. The first term expresses the generating function of the number of customers in the system at the start of the service of the customer who is currently in service. The second term is the generating function of the number of customers that arrived during the past service period of the customer who is currently in service. As $\rho \uparrow 1$, this second term satisfies

$$\lim_{\rho\uparrow 1} \frac{1-\tilde{B}_i\left(\sum_{j=1}^N \lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)}{\left(\sum_{j=1}^N \lambda_j\left(1 - z_{jo}^{(1-\rho)}\right)\right)\mathbb{E}[B_i]} = \lim_{\rho\uparrow 1} \frac{1-\mathbb{E}\left[e^{-\left(\sum_{j=1}^N \lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)B_i}\right]}{\left(\sum_{j=1}^N \lambda_j\left(1 - z_{jo}^{(1-\rho)}\right)\right)\mathbb{E}[B_i]}$$

$$= \lim_{\rho\uparrow 1} \frac{\mathbb{E}\left[B_i e^{-\left(\sum_{j=1}^N \lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)B_i}\right]}{\mathbb{E}[B_i]} = \frac{\mathbb{E}[B_i]}{\mathbb{E}[B_i]} = 1, \tag{6.3}$$

where the second equality follows from l'Hôpital's rule. Equation (6.3) expresses the fact that the scaled vector of number of customers arriving at the different stations during a past service time tends to 0, and hence its generating function tends to 1, as $\rho \uparrow 1$.

Before taking the limit $\rho \uparrow 1$ in (6.2), we first look at $\lim_{\rho\uparrow 1}\left(\tilde{B}_i\left(\sum_{j=1}^N \lambda_j(1 - z_{jo}^{(1-\rho)})\right)\right)^{\frac{1}{1-\rho}}$. As we mentioned earlier, when we scale $\rho \uparrow 1$ we scale the system such that only the arrival rates increase and the service times remain the same. So, we can write $\lambda_j = \rho\hat{\lambda}_j$ where $\hat{\lambda}_j$ is fixed and independent of ρ, for all $j = 1, \ldots, N$. Therefore, we have

$$\lim_{\rho\uparrow 1}\left(\tilde{B}_i\left(\rho\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)\right)\right)^{\frac{1}{1-\rho}}$$

$$= e^{\lim_{\rho\uparrow 1}\frac{\ln\left(\tilde{B}_i\left(\rho\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)\right)\right)}{1-\rho}}$$

$$= e^{\lim_{\rho\uparrow 1}\frac{\left(\rho\sum_{j=1}^{N}\hat{\lambda}_j z_{jo}^{(1-\rho)}\ln z_{jo}+\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)\right)\tilde{B}_i{}'\left(\rho\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)\right)}{-\tilde{B}_i\left(\rho\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)\right)}}$$

$$= e^{\mathbb{E}[B_i]\sum_{j=1}^{N}\hat{\lambda}_j\ln z_{jo}} = e^{\sum_{j=1}^{N}\ln z_{jo}^{\mathbb{E}[B_i]\hat{\lambda}_j}} = \prod_{j=1}^{N}z_{jo}^{\mathbb{E}[B_i]\hat{\lambda}_j}, \tag{6.4}$$

where the second equality follows from l'Hôpital's rule.

Next, we evaluate the following limit, which is related to the denominator of the first term in (6.2)

$$\lim_{\rho\uparrow 1}\frac{z_{iq}^{(1-\rho)}-\tilde{B}_i\left(\sum_{j=1}^{N}\lambda_j\left(1-z_{jo}^{(1-\rho)}\right)\right)}{1-\rho} = \lim_{\rho\uparrow 1}\frac{z_{iq}^{(1-\rho)}-1}{1-\rho}$$

$$+\lim_{\rho\uparrow 1}\frac{1-\mathbb{E}\left[e^{-\rho\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)B_i}\right]}{1-\rho} = \lim_{\rho\uparrow 1}z_{iq}^{(1-\rho)}\ln z_{iq}$$

$$-\lim_{\rho\uparrow 1}\mathbb{E}\left[\left(\rho B_i\sum_{j=1}^{N}\hat{\lambda}_j z_{jo}^{(1-\rho)}\ln z_{jo}+B_i\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)\right)e^{-\rho\sum_{j=1}^{N}\hat{\lambda}_j\left(1-z_{jo}^{(1-\rho)}\right)B_i}\right]$$

$$= \ln z_{iq}-\sum_{j=1}^{N}\hat{\lambda}_j\mathbb{E}[B_i]\ln z_{jo} = \ln\left(z_{iq}\prod_{j=1}^{N}z_{jo}^{-\mathbb{E}[B_i]\hat{\lambda}_j}\right), \tag{6.5}$$

where the first equality uses the fact that $\lambda_j = \rho\hat{\lambda}_j$ and the second equality follows from l'Hôpital's rule.

We know that $\lim_{\rho\uparrow 1}z_{iq}^{(1-\rho)} = 1$. Substituting this along with (6.3), (6.4) and (6.5) in (6.2) while we take $\rho\uparrow 1$, we get

$$\lim_{\rho\uparrow 1}R_{vi}^{(i)}\left(\tilde{z}_q,\tilde{z}_o\right) =$$

$$\lim_{\rho\uparrow 1}\frac{\mathbb{E}\left[z_{iq}^{(1-\rho)Y_i^{(iq)}}\left(\prod_{j=1,j\neq i}^{N}z_{jo}^{(1-\rho)Y_j^{(i)}}\right)z_{io}^{(1-\rho)Y_i^{(io)}}\right]-\mathbb{E}\left[\prod_{j=1}^{N}z_{jo}^{\mathbb{E}[B_i]\hat{\lambda}_j(1-\rho)Y_i^{(iq)}}\left(\prod_{j=1,j\neq i}^{N}z_{jo}^{(1-\rho)Y_j^{(i)}}\right)z_{io}^{(1-\rho)Y_i^{(io)}}\right]}{\mathbb{E}\left[(1-\rho)Y_i^{(iq)}\right]\ln\left(z_{iq}\prod_{j=1}^{N}z_{jo}^{-\mathbb{E}[B_i]\hat{\lambda}_j}\right)}.$$

$$\tag{6.6}$$

Consider the following notation

$$\kappa := \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\left(1-e^{-v_1 G_1}\right)\hat{u}_1^{(1)} = \frac{b^{(2)}}{2b^{(1)}}\frac{\hat{\lambda}_1}{\delta}$$

$$\kappa_1 := \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}e^{-v_1 G_1}\hat{u}_1^{(1)}$$

$$\kappa_i := \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\hat{u}_i^{(1)}, \qquad \forall i = 2,\ldots,N.$$

Using the above notation in (5.1) and substituting it in (6.7) we have,

$$
\lim_{\rho \uparrow 1} R_{vi}^{(1)}\left(\tilde{\underline{z}}_q, \tilde{\underline{z}}_o\right) = \frac{\mathbb{E}\left[z_{1q}^{\kappa\Gamma(\alpha,1)} \prod_{j=1}^{N} z_{jo}^{\kappa_j\Gamma(\alpha,1)}\right] - \mathbb{E}\left[\prod_{j=1}^{N} z_{jo}^{\mathbb{E}[B_1]\hat{\lambda}_j\kappa\Gamma(\alpha,1)} \prod_{j=1}^{N} z_{jo}^{\kappa_j\Gamma(\alpha,1)}\right]}{\mathbb{E}[\kappa\Gamma(\alpha,1)] \ln\left(z_{1q} \prod_{j=1}^{N} z_{jo}^{-\mathbb{E}[B_1]\hat{\lambda}_j}\right)}
$$

$$
= \frac{\mathbb{E}\left[\left(z_{1q}^{\kappa} \prod_{j=1}^{N} z_{jo}^{\kappa_j}\right)^{\Gamma(\alpha,1)}\right] - \mathbb{E}\left[\left(\prod_{j=1}^{N} z_{jo}^{\mathbb{E}[B_1]\hat{\lambda}_j\kappa+\kappa_j}\right)^{\Gamma(\alpha,1)}\right]}{\kappa\alpha \ln\left(z_{1q} \prod_{j=1}^{N} z_{jo}^{-\mathbb{E}[B_1]\hat{\lambda}_j}\right)}.
$$

Now we introduce the following notation to change our generating function into an LST,

$$
s := -\ln z_{1q}
$$
$$
s_i := -\ln z_{io}, \qquad \forall i = 1, \ldots, N.
$$

Then we have that the joint LST of the scaled steady-state joint queue length vector, of the queue of station 1 and the orbits at all the stations, during an arbitrary time in the visit period of station 1 is

$$
\lim_{\rho \uparrow 1} R_{vi}^{(1)}\left(\tilde{\underline{z}}_q, \tilde{\underline{z}}_o\right) = \frac{\mathbb{E}\left[\left(e^{-s\kappa} \prod_{j=1}^{N} e^{-s_j\kappa_j}\right)^{\Gamma(\alpha,1)}\right] - \mathbb{E}\left[\left(\prod_{j=1}^{N} e^{-s_j\left(\mathbb{E}[B_1]\hat{\lambda}_j\kappa+\kappa_j\right)}\right)^{\Gamma(\alpha,1)}\right]}{\alpha\kappa \ln\left(e^{-s} \prod_{j=1}^{N} e^{s_j\mathbb{E}[B_1]\hat{\lambda}_j}\right)}
$$

$$
= \frac{\mathbb{E}\left[e^{-\left(s\kappa + \sum_j s_j\kappa_j\right)\Gamma(\alpha,1)}\right] - \mathbb{E}\left[e^{-\sum_{j=1}^{N} s_j\left(\mathbb{E}[B_1]\hat{\lambda}_j\kappa+\kappa_j\right)\Gamma(\alpha,1)}\right]}{\alpha\kappa\left(-s + \mathbb{E}[B_1]\sum_j \hat{\lambda}_j s_j\right)}
$$

$$
= \frac{\left(\frac{1}{1+s\kappa+\sum_j s_j\kappa_j}\right)^{\alpha} - \left(\frac{1}{1+\sum_{j=1}^{N} s_j\left(\mathbb{E}[B_1]\hat{\lambda}_j\kappa+\kappa_j\right)}\right)^{\alpha}}{\alpha\kappa\left(-s + \mathbb{E}[B_1]\sum_j \hat{\lambda}_j s_j\right)}
$$

$$
= \mathbb{E}\left[e^{-\left(s\kappa + \sum_j s_j\kappa_j + \left(-s\kappa + \sum_j \hat{\lambda}_j\mathbb{E}[B_1]\kappa s_j\right)U\right)\Gamma(\alpha+1,1)}\right],
$$
(6.7)

where U is a standard uniform random variable and the last equality follows from the expression

$$
\mathbb{E}[e^{-(a+bU)\Gamma(\alpha+1,1)}] = \frac{\left(\frac{1}{1+a}\right)^{\alpha} - \left(\frac{1}{1+(a+b)}\right)^{\alpha}}{\alpha b}.
$$

Now we substitute $s = -\ln z_{1q}$ and $s_i = -\ln z_{io}$ back in (6.7) to get for the joint generating function

$$\lim_{\rho \uparrow 1} R_{vi}^{(1)}\left(\tilde{\underline{z}}_q, \tilde{\underline{z}}_o\right) = \mathbb{E}\left[\left(z_{1q}^\kappa \prod_{j=1}^N z_{jo}^{\kappa_j}\left(z_{1q}^{-1} \prod_{j=1}^N z_{jo}^{\hat{\lambda}_j \mathbb{E}[B_1]}\right)^{\kappa U}\right)^{\Gamma(\alpha+1,1)}\right]. \quad (6.8)$$

Let $V_{iq}^{(i)}$ and $V_{io}^{(i)}$ be the number of customers in the queue and orbit of station i, and $V_j^{(i)}$ be the number of customers of type $j \neq i$, at an arbitrary point in time during a visit period of station i, for all $i = 1, \ldots, N$. Then from (6.8) we have,

$$(1-\rho)\begin{pmatrix} V_1^{(1q)} \\ V_1^{(1o)} \\ V_2^{(1)} \\ \vdots \\ V_N^{(1)} \end{pmatrix} \underset{d}{\to} \left[\begin{pmatrix} \kappa \\ \kappa_1 \\ \kappa_2 \\ \vdots \\ \kappa_N \end{pmatrix} + \kappa U \begin{pmatrix} -1 \\ \hat{\lambda}_1 E[B_1] \\ \hat{\lambda}_2 E[B_1] \\ \vdots \\ \hat{\lambda}_N E[B_1] \end{pmatrix}\right]\Gamma(\alpha+1,1), \quad \text{when} \quad \rho \uparrow 1.$$

$$(6.9)$$

Therefore, the scaled steady-state joint queue length vector, in the queue of station 1 and the orbits of all stations, at an arbitrary time during a visit period of station 1, as $\rho \uparrow 1$, satisfies

$$(1-\rho)\begin{pmatrix} V_1^{(1q)} \\ V_1^{(1o)} \\ V_2^{(1)} \\ \vdots \\ V_N^{(1)} \end{pmatrix} \underset{d}{\to} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\left[\begin{pmatrix} (1-e^{-\nu_1 G_1})\hat{u}_1^{(1)} \\ e^{-\nu_1 G_1}\hat{u}_1^{(1)} \\ \hat{u}_2^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix} + U \begin{pmatrix} -\hat{\lambda}_1 \\ \hat{\lambda}_1\hat{\rho}_1 \\ \hat{\lambda}_2\hat{\rho}_1 \\ \vdots \\ \hat{\lambda}_N\hat{\rho}_1 \end{pmatrix}\right]\Gamma(\alpha+1,1), \quad \text{when} \quad \rho \uparrow 1.$$

An intuitive argument for Theorem 4 can be given in the following way. Since, under heavy traffic, the scaled number of customers which are in the queue of station 1 at the start of an arbitrary visit period is gamma distributed, $\kappa\Gamma(\alpha,1)$, also the scaled length of an arbitrary visit period is gamma distributed, $\kappa\mathbb{E}[B_1]\Gamma(\alpha,1)$. Therefore, if we choose an arbitrary time point in a visit period of station 1 the scaled length of that special visit period is distributed as $\kappa\mathbb{E}[B_1]\Gamma(\alpha+1,1)$, where $\kappa = \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}(1-e^{-\nu_1 G_1})\hat{u}_1^{(1)}$. Since this is a special interval which we are looking at, the scaled steady-state joint queue length vector at the start of this visit period satisfies

$$(1-\rho)\begin{pmatrix} \breve{Y}_1^{(1q)} \\ \breve{Y}_1^{(1o)} \\ \breve{Y}_2^{(1)} \\ \vdots \\ \breve{Y}_N^{(1)} \end{pmatrix} \underset{d}{\to} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\begin{pmatrix} (1-e^{-\nu_1 G_1})\hat{u}_1^{(1)} \\ e^{-\nu_1 G_1}\hat{u}_1^{(1)} \\ \hat{u}_2^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix}\Gamma(\alpha+1,1), \quad \text{when} \quad \rho \uparrow 1.$$

$$(6.10)$$

At the arbitrary point in time, we have $\kappa U \Gamma(\alpha + 1, 1)$ customers served, which means that there are $J_j^{(1)} = \sum_{k=1}^{\kappa U \Gamma(\alpha+1,1)} (L_j^{(1)})_k$ new customers of type j arriving during that period. Note that as $\rho \uparrow 1, J_j^{(1)} \to \infty$, therefore the new arrivals during the past service time of the customer in service can be neglected. Using the same arguments as in Theorem 3, we can say that the limiting scaled distribution of the new number of customers of type j at an arbitrary point in time during the visit of station i can be given as

$$\frac{J_j^{(1)}}{U \breve{Y}_1^{(1q)}} \to \lambda_j \mathbb{E}[B_1].$$

Therefore, the scaled steady-state joint queue length vector at an arbitrary point in time during the visit of station 1 as $\rho \uparrow 1$ satisfies

$$(1-\rho) \begin{pmatrix} V_1^{(1q)} \\ V_1^{(1o)} \\ V_2^{(1)} \\ \vdots \\ V_N^{(1)} \end{pmatrix} \xrightarrow{d} \frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \left[\begin{pmatrix} 0 \\ e^{-v_1 G_1} \hat{u}_1^{(1)} \\ \hat{u}_2^{(1)} \\ \vdots \\ \hat{u}_N^{(1)} \end{pmatrix} + (1 - e^{-v_1 G_1}) \hat{u}_1^{(1)} \begin{pmatrix} 1-U \\ U \hat{\lambda}_1 E[B_1] \\ U \hat{\lambda}_2 E[B_1] \\ \vdots \\ U \hat{\lambda}_N E[B_1] \end{pmatrix} \right] \Gamma(\alpha + 1, 1), \quad \text{when} \quad \rho \uparrow 1,$$

which is equivalent to (6.1).

In (6.1), we have given the scaled steady-state joint queue length vector of customers of each type at an arbitrary point in time during the visit period of station 1 when $\rho \uparrow 1$. Using Remark 4, we can extend this to an arbitrary point in time during the visit period of a given station i. This can be written as

$$(1-\rho) \begin{pmatrix} V_1^{(i)} \\ \vdots \\ V_{i-1}^{(i)} \\ V_i^{(iq)} \\ V_i^{(io)} \\ V_{i+1}^{(i)} \\ \vdots \\ V_N^{(i)} \end{pmatrix} \xrightarrow{d} \frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \left[\begin{pmatrix} \hat{u}_1^{(i)} \\ \vdots \\ \hat{u}_{i-1}^{(i)} \\ (1 - e^{-v_i G_i}) \hat{u}_i^{(i)} \\ e^{-v_i G_i} \hat{u}_i^{(i)} \\ \hat{u}_{i+1}^{(i)} \\ \vdots \\ \hat{u}_N^{(i)} \end{pmatrix} + U \begin{pmatrix} \hat{\lambda}_1 \hat{\rho}_i \\ \vdots \\ \hat{\lambda}_{i-1} \hat{\rho}_i \\ -\hat{\lambda}_i \\ \hat{\lambda}_i \hat{\rho}_i \\ \hat{\lambda}_{i+1} \hat{\rho}_i \\ \vdots \\ \hat{\lambda}_N \hat{\rho}_i \end{pmatrix} \right] \Gamma(\alpha + 1, 1), \quad \text{when} \quad \rho \uparrow 1.$$

$$(6.11)$$

Due to the observation that in heavy traffic, the server resides in a visit period for 100% of the time, (6.11) leads to the following theorem.

Theorem 5. *In a cyclic polling system with retrials and glue periods, the scaled steady-state joint queue length vector at an arbitrary time point, with $L^{(iq)}$ and $L^{(io)}$ representing the number in queue and in orbit at station i respectively for all $i = 1, \ldots, N$, satisfies*

$$(1-\rho)\begin{pmatrix} L^{(1q)} \\ \vdots \\ L^{(Nq)} \\ L^{(1o)} \\ \vdots \\ L^{(No)} \end{pmatrix} \xrightarrow{d} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\underline{P}\ \Gamma(\alpha+1,1) \quad \text{when} \quad \rho \uparrow 1, \qquad (6.12)$$

where $\underline{P} = \underline{P}_i$ with probability $\hat{\rho}_i$ and

$$\underline{P}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (1-e^{-v_i G_i})\hat{u}_i^{(i)} \\ 0 \\ \vdots \\ 0 \\ \hat{u}_1^{(i)} \\ \vdots \\ \hat{u}_{i-1}^{(i)} \\ e^{-v_i G_i}\hat{u}_i^{(i)} \\ \hat{u}_{i+1}^{(i)} \\ \vdots \\ \hat{u}_N^{(i)} \end{pmatrix} + U \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -\hat{\lambda}_i \\ 0 \\ \vdots \\ 0 \\ \hat{\lambda}_1\hat{\rho}_i \\ \vdots \\ \hat{\lambda}_{i-1}\hat{\rho}_i \\ \hat{\lambda}_i\hat{\rho}_i \\ \hat{\lambda}_{i+1}\hat{\rho}_i \\ \vdots \\ \hat{\lambda}_N\hat{\rho}_i \end{pmatrix}.$$

Proof. As mentioned at the beginning of this section, when $\hat{\rho} \uparrow 1$ the system is in the visit periods with probability 1. Therefore, the limiting scaled joint queue length distribution at an arbitrary point in time can be given as the limiting scaled joint queue length distribution in the visit period of station i with probability $\hat{\rho}_i$. Now consider that the number of customers of type j, at an arbitrary point in time during the visit period of station i, in queue and orbit respectively is given by $V_j^{(iq)}$ and $V_j^{(io)}$. Then using (6.11), we can write

$$(1-\rho)\begin{pmatrix} V_1^{(iq)} \\ \vdots \\ V_{i-1}^{(iq)} \\ V_i^{(iq)} \\ V_{i+1}^{(iq)} \\ \vdots \\ V_N^{(iq)} \\ V_1^{(io)} \\ \vdots \\ V_{i-1}^{(io)} \\ V_i^{(io)} \\ V_{i+1}^{(io)} \\ \vdots \\ V_N^{(io)} \end{pmatrix} \xrightarrow{d} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \\ (1-e^{-v_i G_i})\hat{u}_i^{(i)} \\ 0 \\ \vdots \\ 0 \\ \hat{u}_1^{(i)} \\ \vdots \\ \hat{u}_{i-1}^{(i)} \\ e^{-v_i G_i}\hat{u}_i^{(i)} \\ \hat{u}_{i+1}^{(i)} \\ \vdots \\ \hat{u}_N^{(i)} \end{pmatrix} + U\begin{pmatrix} 0 \\ \vdots \\ 0 \\ -\hat{\lambda}_i \\ 0 \\ \vdots \\ 0 \\ \hat{\lambda}_1\hat{\rho}_i \\ \vdots \\ \hat{\lambda}_{i-1}\hat{\rho}_i \\ \hat{\lambda}_i\hat{\rho}_i \\ \hat{\lambda}_{i+1}\hat{\rho}_i \\ \vdots \\ \hat{\lambda}_N\hat{\rho}_i \end{pmatrix}\right)\Gamma(\alpha+1,1), \quad \text{when} \quad \rho\uparrow 1.$$

(6.13)

This holds because $V_j^{(io)} = V_j^{(i)}$ and $V_j^{(iq)} = 0$ when $i \neq j$. Therefore, we know that the limiting scaled joint queue length distribution at an arbitrary point in time, with probability $\hat{\rho}_i$, can be given as $\frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\underline{P}_i\Gamma(\alpha+1,1)$. Hence we have

$$(1-\rho)\begin{pmatrix} L^{(1q)} \\ \vdots \\ L^{(Nq)} \\ L^{(1o)} \\ \vdots \\ L^{(No)} \end{pmatrix} \xrightarrow{d} \frac{b^{(2)}}{2b^{(1)}}\frac{1}{\delta}\underline{P}\ \Gamma(\alpha+1,1), \quad \text{when} \quad \rho\uparrow 1, \qquad (6.14)$$

where $\underline{P} = \underline{P}_i$ with probability $\hat{\rho}_i$. □

Remark 7. Note that under heavy-traffic the total scaled workload in the system satisfies the so-called *heavy-traffic averaging principle*. This principle, first found in Refs[6,7] for a specific class of polling models, implies that the workload in each queue is emptied and refilled at a rate that is much faster than the rate at which the total workload is changing. As a consequence, the total workload can be considered constant during the course of a cycle (represented by the gamma distribution), while the workloads in the individual queues fluctuate like a fluid model. It is because of this that the queue length vector in Theorem 5 also features a state-space collapse (cf. Ref[19]): the limiting distribution of the 2N-dimensional scaled queue length vector is governed by just three distributions: the discrete

distribution governing $\underline{P}$, the uniform distribution and the gamma distribution.

Therefore, using Theorems 4 and 5, and the fact that $\delta = \sum_i \mathbb{E}[B_i] \hat{u}_i^{(j)}$ for all $j = 1, \ldots, N$, the scaled workload in the system at an arbitrary point in time is given by

$$(1-\rho) \sum_{i=1}^{N} \mathbb{E}[B_i] \big(L^{(iq)} + L^{(io)}\big) \xrightarrow[d]{} \frac{b^{(2)}}{2b^{(1)}} \Gamma(\alpha + 1, 1), \quad \text{when} \quad \rho \uparrow 1.$$

Since the above equation is independent of everything but the gamma distribution, the workload is the same at any arbitrary point in time during the cycle. Hence the system agrees with the *heavy-traffic averaging principle*. Note that in Theorems 1, 2 and 3, we have found the limiting distribution of the scaled number of customers at embedded time points. Extending the heavy traffic principle along with these theorems, we can say that the scaled workload in an arbitrarily chosen cycle can be given as

$$(1-\rho) \sum_{i=1}^{N} \mathbb{E}[B_i] X_i^{(1)} \xrightarrow[d]{} \frac{b^{(2)}}{2b^{(1)}} \Gamma(\alpha, 1), \quad \text{when} \quad \rho \uparrow 1.$$

We observe that the scaled workload in the two cases, arbitrary point in time and arbitrary cycle, have $\Gamma(\alpha + 1, 1)$ and $\Gamma(\alpha, 1)$ distributions respectively. This is because of a bias introduced in selection of an arbitrary point in time, i.e., an arbitrarily chosen point in time has a higher probability to be in a longer cycle than being in a shorter cycle. This bias does not exist when we arbitrarily choose a cycle.

7. Approximations

In the previous section, we derived the heavy traffic limit of the scaled steady-state joint queue length vector at an arbitrary point in time for the cyclic polling system with retrials and glue periods. We now show that these results are not just valid for the limiting heavy-traffic regime, but can in fact be used to obtain approximations for arbitrary values of the system load. To achieve this, we deploy an interpolation approximation between the heavy traffic result and a light traffic result, similar to what is described in Boon et al.[4], in Section 7.1. Then we give a numerical example in Section 7.2 to illustrate that the error of the approximation is small for arbitrary values of the system load.

7.1. Approximate mean number of customers

Consider the following approximation for the mean number of customers of type i,

$$\mathbb{E}[L_i] \approx \frac{c_0 + \rho c_1}{1 - \rho}. \tag{7.1}$$

The coefficients c_0 and c_1 are chosen in agreement with the light traffic and heavy traffic behavior of $\mathbb{E}[L_i]$. Clearly, when $\rho \downarrow 0$, also $\mathbb{E}[L_i] \downarrow 0$. Hence, we choose $c_0 = 0$. On the other hand, we have $\lim_{\rho \uparrow 1} \mathbb{E}[(1-\rho)L_i] = c_1$. Using (6.12), we get

$$\lim_{\rho \uparrow 1} \mathbb{E}[(1-\rho)L_i] = \lim_{\rho \uparrow 1} \mathbb{E}\left[(1-\rho)\left(L^{(iq)} + L^{(io)}\right)\right]$$

$$= \mathbb{E}\left[\frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \left(\sum_{j=1}^{N} \hat{\rho}_j \left(\hat{u}_i^{(j)} + U\hat{\lambda}_i \hat{\rho}_j\right) - U\hat{\lambda}_i \hat{\rho}_i\right) \Gamma(\alpha + 1, 1)\right]$$

$$= \frac{b^{(2)}}{2b^{(1)}} \frac{1}{\delta} \left(\sum_{j=1}^{N} \hat{\rho}_j \left(\hat{u}_i^{(j)} + \frac{\hat{\lambda}_i \hat{\rho}_j}{2}\right) - \frac{\hat{\lambda}_i \hat{\rho}_i}{2}\right)(\alpha + 1).$$

Hence, we choose $c_1 = \frac{b^{(2)}}{2b^{(1)}} \frac{(\alpha+1)}{\delta} \left(\sum_{j=1}^{N} \hat{\rho}_j(\hat{u}_i^{(j)} + \frac{\hat{\lambda}_i \hat{\rho}_j}{2}) - \frac{\hat{\lambda}_i \hat{\rho}_i}{2}\right)$, and so the final approximation for $\mathbb{E}[L_i]$ becomes, for arbitrary $\rho \in (0,1)$,

$$\mathbb{E}[L_i] \approx \frac{\rho}{(1-\rho)} \frac{b^{(2)}}{2b^{(1)}} \frac{(\alpha+1)}{\delta} \left(\sum_{j=1}^{N} \hat{\rho}_j \left(\hat{u}_i^{(j)} + \frac{\hat{\lambda}_i \hat{\rho}_j}{2}\right) - \frac{\hat{\lambda}_i \hat{\rho}_i}{2}\right). \tag{7.2}$$

7.2. Numerical example

In this section, we will compare the above approximation with exact results. The exact results are obtained using the approach in Ref[2]

Consider a five-station polling system in which the service times are exponentially distributed with mean $\mathbb{E}[B_i] = 1$ for all $i = 1, \ldots, 5$. The arrival processes are Poisson processes with rates $\lambda_1 = \rho \frac{1}{10}, \lambda_2 = \rho \frac{2}{10}, \lambda_3 = \rho \frac{3}{10}, \lambda_4 = \rho \frac{1}{10}, \lambda_5 = \rho \frac{3}{10}$. The switch-over times from station i are exponentially distributed with mean $\mathbb{E}[S_i] = 2, 3, 1, 5, 2$ for stations $i = 1, \ldots, 5$. The durations of the deterministic glue periods are $G_i = 3, 1, 2, 1, 2$, and the exponential retrial rates are $\nu_i = 5, 1, 3, 2, 1$, for stations $i = 1, \ldots, 5$ respectively. We plot the following for $\rho \in (0,1)$ and compare the approximation given in (7.2) with the values obtained using exact analysis.
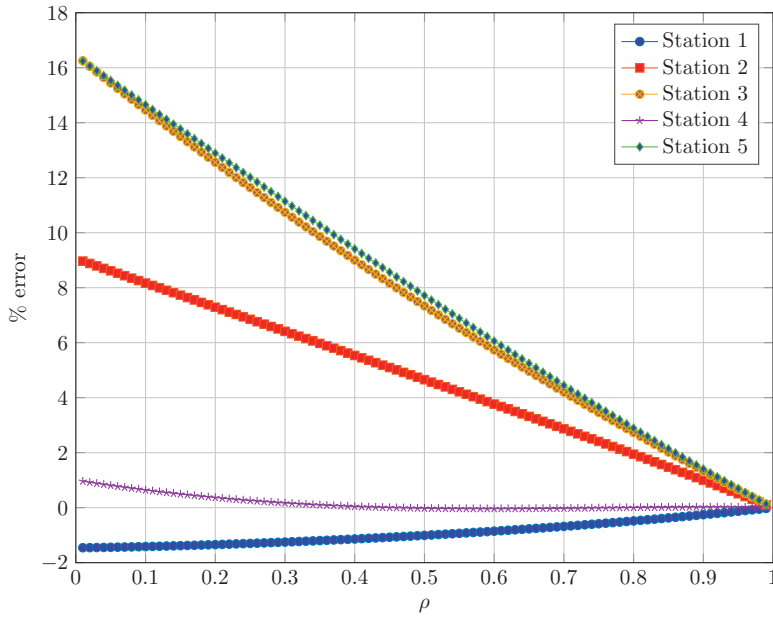
Figure 1. Percentage error for the number of customers in each station.

In Figures 1 and 2, we respectively plot the percentage error calculated as % error $= \frac{\text{Approximate\_value}-\text{Exact\_value}}{\text{Exact\_value}} \times 100$, for the mean number of customers of each type and the total mean number of customers in the system. The error percentage is similar to that predicted in Ref[4]. The error is non-negligible for lower values of ρ, but it decreases quickly as ρ increases. Consequently, for larger values of ρ, the approximation is accurate.
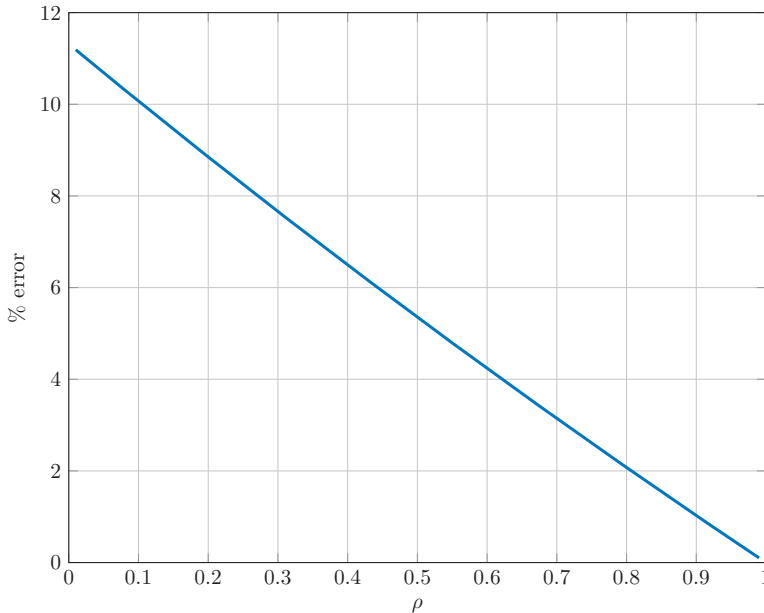

Figure 2. Percentage error for total number of customers.

Based on this, we conclude that the heavy-traffic results as derived in this article are very useful for deriving closed-form approximations for the queue length, especially as the systems under study (e.g., optical systems) typically run under a heavy workload (i.e., a large value of ρ). Nevertheless, to obtain better performance for small values of ρ, the current approximation as presented here can be refined by e.g., computing theoretical values of $\frac{d}{d\rho}\mathbb{E}L_i|_{\rho=0}$ and incorporating that information in (7.1) as explained in Ref[4] Furthermore, approximations for the mean queue length as mentioned here can be extended to approximations for the complete queue length distributions of the polling systems with glue periods and retrials in the spirit of Ref[8]. These extensions, however, are beyond the scope of this article.

Appendix

Proof of Lemma 1. First of all, note that $m_{N,j} = f_{N,j}$ for all $j = 1, ..., N$. Therefore, we have

$$
\mathbf{M_N} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & & 0 \\
0 & 1 & \ddots & \cdots & & 0 \\
\vdots & \ddots & \ddots & \cdots & & 0 \\
0 & \cdots & \cdots & \ddots & & 0 \\
0 & \cdots & \cdots & & 1 & 0 \\
m_{N,1} & m_{N,2} & \cdots & & m_{N,N-1} & m_{N,N}
\end{pmatrix}.
$$

Now using the fact that $m_{N-1,j} = f_{N-1,j} + f_{N-1,N}m_{N,j}$ for all $j \leq N-1$ and furthermore $m_{N-1,N} = f_{N-1,N}m_{N,N}$, we obtain that

$$
\mathbf{M_{N-1}M_N} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & 1 & \ddots & \cdots & \cdots & 0 \\
\vdots & \ddots & \ddots & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & 1 & 0 & 0 \\
m_{N-1,1} & m_{N-1,2} & \cdots & \cdots & m_{N-1,N-1} & m_{N-1,N} \\
m_{N,1} & m_{N,2} & \cdots & \cdots & m_{N,N-1} & m_{N,N}
\end{pmatrix}.
$$

Continuing in this way we obtain

$$
\mathbf{M_1} \cdots \mathbf{M_N} = \begin{pmatrix}
m_{1,1} \cdots m_{1,N} \\
\vdots \ddots \vdots \\
m_{N,1} \cdots m_{N,N}
\end{pmatrix} = \mathbf{M}.
$$

Proof of Lemma 2. First, we look at the normalized right eigenvector $\hat{w}$. Using (3.5), we evaluate the vector $\hat{\mathbf{M}}_{\mathbf{i}}\hat{w}$. Let $(\hat{\mathbf{M}}_{\mathbf{i}}\hat{w})_j$ represent the *jth* element of $\hat{\mathbf{M}}_{\mathbf{i}}\hat{w}$. By a series of simple algebraic manipulations, it follows then that

$$\left(\hat{\mathbf{M}}_{\mathbf{i}}\underline{\hat{w}}\right)_j = \begin{cases} \hat{w}_j, & j \neq i, \\ \dfrac{1}{|\underline{b}|}\sum_{k=1}^{N}\hat{f}_{i,k}\mathbb{E}[B_k], & j = i. \end{cases}$$

However, it also holds that

$$\sum_{k=1}^{N}\hat{f}_{i,k}\mathbb{E}[B_k] = e^{-\nu_i G_i}\mathbb{E}[B_i] + \sum_{k=1}^{N}(1 - e^{-\nu_i G_i})\mathbb{E}[B_i]\hat{\lambda}_k\mathbb{E}[B_k]$$

$$= e^{-\nu_i G_i}\mathbb{E}[B_i] + (1 - e^{-\nu_i G_i})\mathbb{E}[B_i]\sum_{k=1}^{N}\hat{\rho}_k = \mathbb{E}[B_i].$$

Therefore, we conclude that $(\hat{\mathbf{M}}_{\mathbf{i}}\underline{\hat{w}})_i = \hat{w}_i$. This implies that $\underline{\hat{w}}$ is the normalized right eigenvector of $\hat{\mathbf{M}}_{\mathbf{i}}$ for an eigenvalue $\xi = 1$, for all $i = 1, ..., N$. Hence from (3.4), we get the first part of the lemma. Next, we look at the left eigenvector $\underline{\hat{v}}$. Since $\underline{\hat{u}}$ is a multiple of $\underline{\hat{v}}$, it is enough to show that $\underline{\hat{u}}$ is an eigenvector of $\hat{\mathbf{M}}$. Define

$$\underline{u}^{(i)} = \begin{pmatrix} u_1^{(i)} \\ \vdots \\ u_N^{(i)} \end{pmatrix}, \qquad \text{where} \qquad u_j^{(i)} = \begin{cases} \lambda_j\left[\dfrac{e^{-\nu_j G_j}}{1 - e^{-\nu_j G_j}} + \sum_{k=j}^{N}\rho_k + \sum_{k=1}^{i-1}\rho_k\right], & i \leq j, \\ \lambda_j\left[\dfrac{e^{-\nu_j G_j}}{1 - e^{-\nu_j G_j}} + \sum_{k=j}^{i-1}\rho_k\right], & i > j. \end{cases}$$

$$(A.1)$$

Note that $u_j^{(1)} = u_j^{(N+1)} = u_j$, for all $j = 1, ..., N$, and hence, $\underline{u}^{(1)} = \underline{u}^{(N+1)} = \underline{u}$. Furthermore, we have

$$\underline{\hat{u}}^{(1)^T}\hat{\mathbf{M}}_1 = \begin{pmatrix} \hat{u}_1\hat{f}_{1,1} \\ \hat{u}_1\hat{f}_{1,2} + \hat{u}_2 \\ \vdots \\ \hat{u}_1\hat{f}_{1,N} + \hat{u}_N \end{pmatrix}^T = \begin{pmatrix} \hat{\lambda}_1\frac{e^{-\nu_1 G_1}}{1 - e^{-\nu_1 G_1}} + \hat{\lambda}_1\hat{\rho}_1 \\ \hat{u}_2 + \hat{\lambda}_2\hat{\rho}_1 \\ \vdots \\ \hat{u}_N + \hat{\lambda}_N\hat{\rho}_1 \end{pmatrix}^T = \begin{pmatrix} \hat{u}_1^{(2)} \\ \hat{u}_2^{(2)} \\ \vdots \\ \hat{u}_N^{(2)} \end{pmatrix}^T = \underline{\hat{u}}^{(2)^T},$$

and, in a similar way, for all $i = 1, ..., N$,

$$\underline{\hat{u}}^{(i)^T}\hat{\mathbf{M}}_{\mathbf{i}} = \underline{\hat{u}}^{(i+1)^T}. \qquad (A.2)$$

Therefore, we have

$$\underline{\hat{u}}^T\mathbf{M} = \underline{\hat{u}}^{(1)^T}\hat{\mathbf{M}}_1 \cdots \hat{\mathbf{M}}_{\mathbf{N}} = \underline{\hat{u}}^{(N+1)^T} = \underline{\hat{u}}^T.$$

Hence $\underline{\hat{u}}$ and $\underline{\hat{v}}$ are the left eigenvectors of $\hat{\mathbf{M}}$, for eigenvalue $\xi = 1$. $\qquad\square$

Proof of Lemma 3. Since the maximal eigenvalue ξ of $\mathbf{M}$ is a simple eigenvalue and furthermore $\mathbf{M}$ is continuous in ρ, Theorem 5 of Lancaster [11] states that

$$\left.\frac{d\xi}{d\rho}\right|_{\rho=1} = \frac{\hat{\underline{v}}^T\hat{\mathbf{M}}'\hat{\underline{w}}}{\hat{\underline{v}}^T\hat{\underline{w}}}, \qquad (A.3)$$

where $\hat{\mathbf{M}}'$ is the element wise derivative of $\mathbf{M}$ with respect to ρ evaluated at $\rho = 1$. Let $\mathbf{U_i} = (\prod_{k=1}^{i-1}\mathbf{M_k})\mathbf{M_i'}(\prod_{k=i+1}^{N}\mathbf{M_k})$. Then due to (3.4) we can write $\mathbf{M'} = \sum_{i=1}^{N}\mathbf{U_i}$. From (3.5) we can see that

$$
\mathbf{M_i'} = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \cdots & \ddots & \vdots \\ \dfrac{df_{i,1}}{d\rho} & \cdots & \dfrac{df_{i,i}}{d\rho} & \cdots & \dfrac{df_{i,N}}{d\rho} \\ \vdots & \ddots & \cdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}
$$

$$
= \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ (1-e^{-\nu_i G_i})\mathbb{E}[B_i]\dfrac{d\lambda_1}{d\rho} & \cdots & (1-e^{-\nu_i G_i})\mathbb{E}[B_i]\dfrac{d\lambda_N}{d\rho} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}
$$

$$
= \begin{pmatrix} 0 \\ \vdots \\ (1-e^{-\nu_i G_i})\mathbb{E}[B_i] \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} \dfrac{d\lambda_1}{d\rho} & \cdots & \dfrac{d\lambda_N}{d\rho} \end{pmatrix}. \tag{A.4}
$$

From the definition of ρ, we know that $\sum_{i=1}^{N} \mathbb{E}[B_i]\frac{d\lambda_i}{d\rho} = 1$, and hence

$$
\begin{pmatrix} \dfrac{d\lambda_1}{d\rho} & \cdots & \dfrac{d\lambda_N}{d\rho} \end{pmatrix} \underline{\hat{w}} = |\underline{b}|^{-1}. \tag{A.5}
$$

Since $\underline{\hat{w}}$ is the normalized right eigenvector of any $\hat{\mathbf{M}}_\mathbf{i}$ for eigenvalue $\xi = 1$, we have

$$
\prod_{k=i+1}^{N} \hat{\mathbf{M}}_\mathbf{k}\underline{\hat{w}} = \underline{\hat{w}}. \tag{A.6}
$$

Using (A.4), (A.5) and (A.6) we get

$$
\hat{\mathbf{M}}_\mathbf{i}' \prod_{k=i+1}^{N} \hat{\mathbf{M}}_\mathbf{k}\underline{\hat{w}} = \frac{1}{|\underline{b}|} \begin{pmatrix} 0 \\ \vdots \\ (1-e^{-\nu_i G_i})\mathbb{E}[B_i] \\ \vdots \\ 0 \end{pmatrix}. \tag{A.7}
$$

From (A.2) and (A.7) we get

$$
\underline{\hat{u}}^T \hat{\mathbf{U}}_\mathbf{i}\underline{\hat{w}} = \left(\underline{\hat{u}}^T\prod_{k=1}^{i-1}\hat{\mathbf{M}}_\mathbf{k} \right)\hat{\mathbf{M}}_\mathbf{i}'\left(\prod_{k=i+1}^{N}\hat{\mathbf{M}}_\mathbf{k}\underline{\hat{w}} \right) = \frac{1}{|\underline{b}|} \begin{pmatrix} \hat{u}_1^{(i)} \\ \vdots \\ \hat{u}_{i-1}^{(i)} \\ \hat{u}_i^{(i)} \\ \hat{u}_{i+1}^{(i)} \\ \vdots \\ \hat{u}_N^{(i)} \end{pmatrix}^T \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (1-e^{-\nu_i G_i})\mathbb{E}[B_i] \\ 0 \\ \vdots \\ 0 \end{pmatrix}
$$

$$= \frac{\hat{u}_i^{(i)}(1-e^{-\nu_i G_i})\mathbb{E}[B_i]}{|\underline{b}|} = \frac{\hat{\rho}_i}{|\underline{b}|}, \tag{A.8}$$

where the last equality follows from the fact that $\hat{u}_i^{(i)} = \hat{\lambda}_i/(1-e^{-\nu_i G_i})$, see (A.1). Multiplying both sides of (A.8) with $|\underline{b}|/\delta$ and summing it over all $i = 1, ..., N$, we get that

$$\hat{\underline{v}}^T \hat{\mathbf{M}}' \hat{\underline{w}} = \sum_{i=1}^{N} \frac{|\underline{b}|}{\delta} \hat{\underline{u}}^T \hat{\mathbf{U}}_i \hat{\underline{w}} = \sum_{i=1}^{N} \frac{\hat{\rho}_i}{\delta} = \frac{1}{\delta}. \tag{A.9}$$

Since $\hat{\underline{v}}^T \hat{\underline{w}} = 1$, we obtain from (A.3) and (A.9) that $\xi'(1) = \frac{1}{\delta}$. $\qquad\square$

Proof of Lemma 4. We know that

$$h_i(\underline{z}) = f_i\big(z_1, ..., z_i, h_{i+1}(\underline{z}), ..., h_N(\underline{z})\big)$$
$$= (1-e^{-\nu_i G_i})\beta_i\big(z_1, ..., z_i, h_{i+1}(\underline{z}), ..., h_N(\underline{z})\big) + e^{-\nu_i G_i}z_i$$
$$= (1-e^{-\nu_i G_i})\mathbb{E}\left[e^{-B_i\left(\sum_{c=1}^{i}(1-z_c)\lambda_c + \sum_{c=i+1}^{N}(1-h_c(\underline{z}))\lambda_c\right)}\right] + e^{-\nu_i G_i}z_i.$$

From this it follows that

$$\frac{\partial h_i(\underline{z})}{\partial z_k} = (1-e^{-\nu_i G_i})\mathbb{E}\left[B_i\left(\lambda_k 1[k \le i] + \sum_{c=i+1}^{N}\lambda_c\frac{\partial h_c(\underline{z})}{\partial z_k}\right)e^{-B_i\left(\sum_{c=1}^{i}(1-z_c)\lambda_c + \sum_{c=i+1}^{N}(1-h_c(\underline{z}))\lambda_c\right)}\right] + e^{-\nu_i G_i}1[k = i],$$

and $\quad \dfrac{\partial^2 h_i(\underline{z})}{\partial z_j \partial z_k} = (1-e^{-\nu_i G_i})\mathbb{E}\left[\left(B_i^2\left(\lambda_k 1[k \le i] + \sum_{c=i+1}^{N}\lambda_c\frac{\partial h_c(\underline{z})}{\partial z_k}\right)\left(\lambda_j 1[j \le i] + \sum_{c=i+1}^{N}\lambda_c\frac{\partial h_c(\underline{z})}{\partial z_j}\right)\right.\right.$

$$\left.\left.+ B_i\sum_{c=i+1}^{N}\lambda_c\frac{\partial^2 h_c(\underline{z})}{\partial z_j \partial z_k}\right)e^{-B_i\left(\sum_{c=1}^{i}(1-z_c)\lambda_c + \sum_{c=i+1}^{N}(1-h_c(\underline{z}))\lambda_c\right)}\right],$$

$$\tag{A.10}$$

where $1[E] = 1$, when the event E is true and otherwise $1[E] = 0$. Because

$$\left.\frac{\partial^2 h_i(\underline{z})}{\partial z_j \partial z_k}\right|_{\underline{z}=\underline{1}} = k_{j,k}^{(i)}$$

and

$$\left.(1-e^{-\nu_i G_i})\mathbb{E}[B_i]\left(\lambda_k 1[k \le i] + \sum_{c=i+1}^{N}\lambda_c\frac{\delta h_c}{\delta z_k}(\underline{z})\right)\right|_{\underline{z}=\underline{1}} = m_{i,k}-1[k = i]e^{-\nu_i G_i},$$

we have

$$k_{j,k}^{(i)} = \frac{\mathbb{E}[B_i^2]}{\mathbb{E}[B_i]^2(1-e^{-\nu_i G_i})}\left(m_{i,j}-1[j = i]e^{-\nu_i G_i}\right)\left(m_{i,k}-1[k = i]e^{-\nu_i G_i}\right) + (1-e^{-\nu_i G_i})\mathbb{E}[B_i]\sum_{c=i+1}^{N}\lambda_c k_{j,k}^{(c)}$$

$$= \frac{\mathbb{E}[B_i^2]}{\mathbb{E}[B_i]^2(1-e^{-\nu_i G_i})}\left(m_{i,j}m_{i,k}-\left(1[j = i]m_{i,k} + 1[k = i]m_{i,j}\right)e^{-\nu_i G_i} + 1[i = j = k]e^{-2\nu_i G_i}\right)$$

$$+ (1-e^{-\nu_i G_i})\mathbb{E}[B_i]\sum_{c=i+1}^{N}\lambda_c k_{j,k}^{(c)}. \tag{A.11}$$

Let 1_i be an $N \times N$ matrix, where the element in the i-th row and the i-th column equals one, and all N^2-1 other entries read zero. Then, based on (A.11), we can write

$$K^{(i)} = \frac{\mathbb{E}\left[B_i^2\right]}{\mathbb{E}[B_i]^2(1 - \mathrm{e}^{-\nu_i G_i})} \left[\begin{pmatrix} m_{i,1}m_{i,1} \cdots m_{i,1}m_{i,N} \\ \vdots \ddots \vdots \\ m_{i,N}m_{i,1} \cdots m_{i,N}m_{i,N} \end{pmatrix} - \mathrm{e}^{-\nu_i G_i} \begin{pmatrix} 0 & \cdots & m_{i,1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{i,i-1} & \cdots & 0 \\ m_{i,1} & \cdots & 2m_{i,i} & \cdots & m_{i,N} \\ 0 & \cdots & m_{i,i+1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{i,N} & \cdots & 0 \end{pmatrix} + \mathrm{e}^{-2\nu_i G_i} 1_i \right]$$

$$+ (1 - \mathrm{e}^{-\nu_i G_i})\mathbb{E}[B_i] \sum_{c=i+1}^{N} \lambda_c K^{(c)}$$

$$= \frac{\mathbb{E}\left[B_i^2\right]}{\mathbb{E}[B_i]^2(1 - \mathrm{e}^{-\nu_i G_i})} \left[\begin{pmatrix} m_{i,1} \\ \vdots \\ m_{i,N} \end{pmatrix} \begin{pmatrix} m_{i,1} \cdots m_{i,N} \end{pmatrix} - \mathrm{e}^{-\nu_i G_i} \begin{pmatrix} 0 & \cdots & m_{i,1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{i,i-1} & \cdots & 0 \\ m_{i,1} & \cdots & 2m_{i,i} & \cdots & m_{i,N} \\ 0 & \cdots & m_{i,i+1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{i,N} & \cdots & 0 \end{pmatrix} + \mathrm{e}^{-2\nu_i G_i} 1_i \right]$$

$$+ (1 - \mathrm{e}^{-\nu_i G_i})\mathbb{E}[B_i] \sum_{c=i+1}^{N} \lambda_c K^{(c)}.$$

This leads to

$$\underline{w}^T K^{(i)} \underline{w} = \frac{\mathbb{E}\left[B_i^2\right]}{\mathbb{E}[B_i]^2(1 - \mathrm{e}^{-\nu_i G_i})}$$

$$\left[\underline{w}^T \begin{pmatrix} m_{i,1} \\ \vdots \\ m_{i,N} \end{pmatrix} \begin{pmatrix} m_{i,1} \cdots m_{i,N} \end{pmatrix} \underline{w} - \mathrm{e}^{-\nu_i G_i} \underline{w}^T \begin{pmatrix} 0 & \cdots & m_{i,1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{i,i-1} & \cdots & 0 \\ m_{i,1} & \cdots & 2m_{i,i} & \cdots & m_{i,N} \\ 0 & \cdots & m_{i,i+1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{i,N} & \cdots & 0 \end{pmatrix} \underline{w} + \mathrm{e}^{-2\nu_i G_i} \underline{w}^T 1_i \underline{w} \right]$$

$$+ (1 - \mathrm{e}^{-\nu_i G_i})\mathbb{E}[B_i] \sum_{c=i+1}^{N} \lambda_c \underline{w}^T K^{(c)} \underline{w}. \tag{A.12}$$

Note that from the definition of $\underline{\hat{w}}$, we have that

$$\underline{\hat{w}}^T \begin{pmatrix} \hat{m}_{i,1} \\ \vdots \\ \hat{m}_{i,N} \end{pmatrix} = \begin{pmatrix} \hat{m}_{i,1} \cdots \hat{m}_{i,N} \end{pmatrix} \underline{\hat{w}} = \frac{\mathbb{E}[B_i]}{|\underline{b}|}. \tag{A.13}$$

Now we evaluate

$$
\underline{\hat{w}}^T \begin{pmatrix} 0 & \cdots & \hat{m}_{i,1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{m}_{i,i-1} & \cdots & 0 \\ \hat{m}_{i,1} & \cdots & 2\hat{m}_{i,i} & \cdots & \hat{m}_{i,N} \\ 0 & \cdots & \hat{m}_{i,i+1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{m}_{i,N} & \cdots & 0 \end{pmatrix} \underline{\hat{w}} = \frac{1}{|\underline{b}|} \begin{pmatrix} \hat{m}_{i,1}\mathbb{E}[B_i] \\ \vdots \\ \hat{m}_{i,i-1}\mathbb{E}[B_i] \\ \hat{m}_{i,i}\mathbb{E}[B_i] + \sum_{j=1}^{N}\hat{m}_{i,j}\mathbb{E}[B_j] \\ \hat{m}_{i,i+1}\mathbb{E}[B_i] \\ \vdots \\ \hat{m}_{i,N}\mathbb{E}[B_i] \end{pmatrix}^T \underline{\hat{w}}
$$

$$
= \frac{\mathbb{E}[B_i]\sum_{j=1}^{N}\hat{m}_{i,j}\mathbb{E}[B_j] + \mathbb{E}[B_i]\sum_{j=1}^{N}\hat{m}_{i,j}\mathbb{E}[B_j]}{|\underline{b}|^2} = \frac{2\mathbb{E}[B_i]^2}{|\underline{b}|^2}. \tag{A.14}
$$

Evaluating (A.12) for $\rho \uparrow 1$, and substituting (A.13) and (A.14) in it, we have

$$
\underline{\hat{w}}^T \hat{K}^{(i)} \underline{\hat{w}} = \frac{\mathbb{E}[B_i^2]}{\mathbb{E}[B_i]^2(1-e^{-\nu_i G_i})}\left[\frac{\mathbb{E}[B_i]^2}{|\underline{b}|^2} - 2e^{-\nu_i G_i}\frac{\mathbb{E}[B_i]^2}{|\underline{b}|^2} + e^{-2\nu_i G_i}\frac{\mathbb{E}[B_i]^2}{|\underline{b}|^2}\right]
$$

$$
+(1-e^{-\nu_i G_i})\mathbb{E}[B_i]\sum_{c=i+1}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}} = (1-e^{-\nu_i G_i})\frac{\mathbb{E}[B_i^2]}{|\underline{b}|^2}
$$

$$
+(1-e^{-\nu_i G_i})\mathbb{E}[B_i]\sum_{c=i+1}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}} = (1-e^{-\nu_i G_i})\left(\frac{\mathbb{E}[B_i^2]}{|\underline{b}|^2} + \mathbb{E}[B_i]\sum_{c=i+1}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}\right). \tag{A.15}
$$

Multiplying both sides of (A.15) with $\hat{v}_i$ and evaluating it for $i=1$ we get

$$
\hat{v}_1\underline{\hat{w}}^T \hat{K}^{(1)} \underline{\hat{w}} = \frac{|\underline{b}|\hat{\lambda}_1}{\delta}\left(\frac{\mathbb{E}[B_1^2]}{|\underline{b}|^2} + \mathbb{E}[B_1]\sum_{c=2}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}\right) = \frac{\hat{\lambda}_1\mathbb{E}[B_1^2]}{\delta|\underline{b}|}
$$

$$
+\frac{|\underline{b}|\hat{\rho}_1\hat{\lambda}_2(1-e^{-\nu_2 G_2})}{\delta}\left(\frac{\mathbb{E}[B_2^2]}{|\underline{b}|^2} + \sum_{c=3}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}\right) + |\underline{b}|\frac{\hat{\rho}_1}{\delta}\sum_{c=3}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}, \tag{A.16}
$$

where for the second equality we again used (A.15), but now for $i=2$, to substitute $\underline{\hat{w}}^T\hat{K}^{(2)}\underline{\hat{w}}$. Multiplying both sides of (A.15) with $\hat{v}_i$ and evaluating it for $i=2$ we get

$$
\hat{v}_2\underline{\hat{w}}^T \hat{K}^{(2)} \underline{\hat{w}} = \frac{|\underline{b}|\hat{\lambda}_2}{\delta}\left(e^{-\nu_2 G_2} + (1-e^{-\nu_2 G_2})\sum_{j=2}^{N}\hat{\rho}_j\right)\left(\frac{\mathbb{E}[B_2^2]}{|\underline{b}|^2} + \mathbb{E}[B_2]\sum_{c=3}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}\right)
$$

$$
= \frac{|\underline{b}|\hat{\lambda}_2}{\delta}\left(e^{-\nu_2 G_2} + (1-e^{-\nu_2 G_2})(1-\hat{\rho}_1)\right)\left(\frac{\mathbb{E}[B_2^2]}{|\underline{b}|^2} + \mathbb{E}[B_2]\sum_{c=3}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}\right)
$$

$$
= \frac{|\underline{b}|\hat{\lambda}_2 - |\underline{b}|\hat{\rho}_1\hat{\lambda}_2(1-e^{-\nu_2 G_2})}{\delta}\left(\frac{\mathbb{E}[B_2^2]}{|\underline{b}|^2} + \mathbb{E}[B_2]\sum_{c=3}^{N}\hat{\lambda}_c\underline{\hat{w}}^T\hat{K}^{(c)}\underline{\hat{w}}\right). \tag{A.17}
$$

Summing (A.16) and (A.17) we get

$$\sum_{j=1}^{2} \hat{v}_j \hat{\underline{w}}^T \hat{K}^{(j)} \hat{\underline{w}} = \sum_{j=1}^{2} \frac{\hat{\lambda}_j}{\delta} \frac{\mathbb{E}\left[B_j^2\right]}{|\underline{b}|} + \frac{|\underline{b}|}{\delta} \left(\sum_{j=1}^{2} \hat{\rho}_j \right) \sum_{c=3}^{N} \hat{\lambda}_c \hat{\underline{w}}^T \hat{K}^{(c)} \hat{\underline{w}}.$$

By repeating the above procedure, we end up with

$$\sum_{j=1}^{N} \hat{v}_j \hat{\underline{w}}^T \hat{K}^{(j)} \hat{\underline{w}} = \sum_{j=1}^{N} \frac{\hat{\lambda}_j}{\delta} \frac{\mathbb{E}\left[B_j^2\right]}{|\underline{b}|} = \frac{1}{\delta|\underline{b}|} \frac{b^{(2)}}{b^{(1)}}.$$

Therefore we have

$$A := \frac{1}{2} \sum_{j=1}^{N} \hat{v}_j \hat{\underline{w}}^T \hat{K}^{(j)} \hat{\underline{w}} = \frac{1}{2\delta|\underline{b}|} \frac{b^{(2)}}{b^{(1)}}.$$

□

Proof of Lemma 5. Multiplying both sides of (3.6) with $\mathbb{E}[B_i]$ and summing it over all i gives

$$\sum_{i=1}^{N} g_i \mathbb{E}[B_i] = \sum_{k=1}^{N} \lambda_k \left(\sum_{i=1}^{N} m_{k,i} \mathbb{E}[B_i] \right) \left(\sum_{j=1}^{k-1} \left(G_j + \mathbb{E}[S_j] \right) \left(1 - e^{-\nu_k G_k} \right) + G_k \right)$$

$$+ \sum_{i=1}^{N} \rho_i \left(\sum_{j=1}^{i-1} \left(G_j + \mathbb{E}[S_j] \right) e^{-\nu_i G_i} + \sum_{j=i}^{N} \mathbb{E}[S_j] + \sum_{j=i+1}^{N} G_j \right).$$

Since $\hat{w}$ is an eigenvector of $\hat{\mathbf{M}}_\mathbf{k}$, we have $\sum_{i=1}^{N} \hat{m}_{k,i} \mathbb{E}[B_i] = \mathbb{E}[B_k]$. Hence, taking $\rho \uparrow 1$, we get

$$\sum_{i=1}^{N} \hat{g}_i \mathbb{E}[B_i] = \sum_{i=1}^{N} \hat{\rho}_i \left(\sum_{j=1}^{i-1} \left(G_j + \mathbb{E}[S_j] \right) \left(1 - e^{-\nu_i G_i} \right) + G_i + \sum_{j=1}^{i-1} \left(G_j + \mathbb{E}[S_j] \right) e^{-\nu_i G_i} + \sum_{j=i}^{N} \mathbb{E}[S_j] + \sum_{j=i+1}^{N} G_j \right)$$

$$= \sum_{i=1}^{N} \hat{\rho}_i \sum_{j=1}^{N} \left(\mathbb{E}[S_j] + G_j \right) = \sum_{i=1}^{N} \left(\mathbb{E}[S_i] + G_i \right). \tag{A.18}$$

Substituting $\hat{\underline{w}} = |\underline{b}|^{-1} (\mathbb{E}[B_1] \cdots \mathbb{E}[B_N])^T$ and $A = \frac{1}{2\delta|\underline{b}|} \frac{b^{(2)}}{b^{(1)}}$ in (4.1) and using (A.18) will give that $\alpha = 2r\delta \frac{b^{(1)}}{b^{(2)}}$. □

Proof of Lemma 6. First, we state (1.1) and (1.2) of Ref[18] in our notation.

For some $a_1 > 0, a_2 > 0, a_3 < \infty$ and a positive integer U, (1.1) in Ref[18] entails the following conditions:

(i) $\{M^U\}_{i,j} \geq a_1, \quad \forall i, j = 1, 2, ..., N.$

(ii) $\sum_{i,j,k} k_{j,k}^{(i)} \geq a_2, \quad \forall i, j, k = 1, 2, ..., N.$

(iii) $\sum_{i,j,k,l} \frac{\partial^3 h_i(\underline{z})}{\partial z_j \partial z_k \partial z_l} \big|_{\underline{z}=\underline{1}} \leq a_3 \quad \forall i, j, k, l = 1, 2, ..., N.$

Furthermore, for some $a_4 > 0$ and $a_5 < \infty$, (1.2) in [18] entails the following conditions:

(iv) $g(\underline{1}) = 1.$

(v) $\sum_{i=1}^{N} g_i \geq a_4$.

(vi) $\sum_{i,j} \frac{\partial^2 g(\underline{z})}{\partial z_j \partial z_k}\big|_{\underline{z}=\underline{1}} \leq a_5 \quad \forall i,j = 1,2,...,N.$

Now we prove that each of the above statements hold for our model.

First, considering Eqs. (3.2) and (3.3), and assuming that arrival rates and service times are positive (which is the case in a non-trivial model), we conclude that $m_{i,j} > 0$ for all $i,j = 1,2,...,N$. Hence, (i) holds for $U=1$ and a_1 close enough to zero.

Condition (ii) implies that the sum of second order joint moments of the number of children of type j and k produced by a customer of type i should be positive. As numbers of children cannot become negative, all these joint moments are clearly non-negative. This can also be seen from (A.10). Furthermore, when $i = N$, (A.10) reveals that

$$k_{j,k}^{(N)} = \frac{\partial^2 h_N(\underline{z})}{\partial z_j \partial z_k}\bigg|_{\underline{z}=\underline{1}} = (1-e^{-\nu_N G_N})\mathbb{E}\left[B_N^2 \lambda_j \lambda_k e^{-B_N\left(\sum_{c=1}^{N}(1-z_c)\lambda_c\right)}\right]\bigg|_{\underline{z}=\underline{1}} = \lambda_j \lambda_k (1-e^{-\nu_N G_N})\mathbb{E}\left[B_N^2\right].$$

As each of the terms of the above equations are positive in a non-trivial model, so is $k_{j,k}^{(N)}$. Since all the values of $k_{j,k}^{(i)}$ are non-negative and at least one is positive the value of $\sum_{i,j,k} k_{j,k}^{(i)}$ is positive, and hence (ii) holds for a_2 close to zero.

For (iii), note that the arrival processes in our system are independent Poisson processes with finite positive rate and the third moments of the service times are finite. This implies the third-order joint moments are finite and hence their sum is finite. So (iii) holds for some finite value of a_3.

Condition (iv) holds as $g(\underline{z})$ is a generating function of a vector of finite random variables.

Next, it is easy to verify that each term in (3.6) is positive for a non-trivial model. Hence, so are $g_1, ..., g_N$. Therefore, we conclude that (v) holds for a_4 close enough to zero.

The arrival processes in our system are independent Poisson processes with finite positive rate. Further, the second moments of the switchover times are finite, and the glue periods are finite and deterministic. It implies that the second moments of the non-visit times are finite and hence the second-order joint moments are finite and therefore their sum is finite. So (vi) is satisfied for some finite value of a_5.   □

Acknowledgements

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

of the research of the second author was performed while he was affiliated with Leiden University.

References

[1]   Abidini, M.A.; Boxma, O.J.; Kim, B.; Kim, J.; Resing, J.A.C. Performance analysis of polling systems with retrials and glue periods. *Queueing Syst.* 2017, 87, 293–324.

[2]   Abidini, M.A.; Boxma, O.J.; Resing, J.A.C. Analysis and optimization of vacation and polling models with retrials. *Perform. Eval.* 2016, 98, 52–69.

[3]   Boon, M.A.A.; van der Mei, R.D.; Winands, E.M.M. Applications of polling systems. *SORMS* 2011, 16, 67–82.

[4]   Boon, M.A.A.; Winands, E.M.M.; Adan, I.J.B.F.; van Wijk, A.C.C. Closed-form waiting time approximations for polling systems. *Perform. Eval.* 2011, 68, 290–306.

[5]   Boxma, O.J.; Resing, J.A.C. Vacation and polling models with retrials. In *11th European Workshop on Performance Engineering (EPEW 11), Florence, Italy September 2014*, 2014.

[6]   Coffman, E.G.; Puhalskii, A.A.; Reiman, M.I. Polling systems with zero switch-over times: A heavy traffic principle. *Ann. Appl. Probab.* 1995, 5, 681–719.

[7]   Coffman, E.G.; Puhalskii, A.A.; Reiman, M.I. Polling systems in heavy traffic: A Bessel process limit. *Math. Oper. Res.* 1998, 23, 257–304.

[8]   Dorsman, J.L.; van der Mei, R.D.; Winands, E.M.M. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *Stoch. Models* 2011, 27, 318–332.

[9]   Grimmett, G.R.; Stirzaker, D.R. *Probability and Random Processes*; Second Edition, Clarendon Press: Oxford, 1992.

[10] Konheim, A.G.; Levy, H.; Srinivasan, M.M. Descendant set: An efficient approach for the analysis of polling systems. *IEEE Trans. Commun.* 1994, 42, 1245–1253.

[11] Lancaster, P. On eigenvalues of matrices dependent on a parameter. *Numer. Math.* 1964, 6, 377–387.

[12] Levy, H. Binomial-gated service: A method for effective operation and optimization of polling systems. *IEEE Trans. Commun.* 1991, 39, 1341–1350.

[13] Levy, H.; Sidi, M. Polling models: Applications, modeling and optimization. *IEEE Trans. Commun.* 1990, 38, 1750–1760.

[14] van der Mei, R.D. Distribution of the delay in polling systems in heavy traffic. *Perform. Evaluation* 1999, 38, 133–148.

[15] van der Mei, R.D. Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Syst.* 2007, 57, 29–46.

[16] Okawachi, Y.; Bigelow, M.S.; Sharping, J.E.; Zhu, Z.; Schweinsberg, A.; Gauthier, D.J.; Boyd, R. W.; Gaeta, A.L. Tunable all-optical delays via Brillouin slow light in an optical fiber. *Phys. Rev. Lett.* 2005, 94, 153902.

[17] Olsen, T.L.; van der Mei, R.D. Polling systems with periodic server routing in heavy traffic: distribution of the delay. *J. Appl. Probab.* 2003, 40, 305–326.

[18] Quine, M.P. The multitype Galton-Watson process with ρ near 1. *Adv. Appl. Probab.* 1972, 4, 429–452.

[19] Reiman, M.I. Some diffusion approximations with state space collapse. *Model. Perform. Eval. Methodol.* 1984, 60, 209–240.

[20] Resing, J.A.C. Polling systems and multitype branching processes. *Queueing Syst.* 1993, 13, 409–426.

[21]   Takagi, H. Application of polling models to computer networks. *Comp. Netw. ISDN Syst.* 1991, 22, 193–211.

[22]   Takagi, H. Queueing analysis of polling models: progress in 1990-1994. *Front. Queueing Model. Meth. Prob.* 1997, 119–146.

[23]   Takagi, H. Analysis and application of polling models. *Perform. Eval.: Origins and Directions (Lect. Notes Comput. Sc)*, 2000, 1769, 424–442.

[24]   Vishnevskii, V.M.; Semenova, O.V. Mathematical methods to study the polling systems. *Autom. Remote Control* 2006, 67, 173–220.

[25]   Williams, D. 1991. *Probability with Martingales*; Cambridge University Press.

[26]   Zedek, M. Continuity and location of zeros of linear combinations of polynomials. *Proc. Amer. Math. Soc.* 1965, 16, 78–84.