Adv. Appl. Prob. **46,** 76–101 (2014)
Printed in Northern Ireland

RANDOM FLUID LIMIT OF AN OVERLOADED POLLING MODEL

MARIA REMEROVA,* *CWI*
SERGEY FOSS,** *Heriot-Watt University and Sobolev Institute of Mathematics*
BERT ZWART,* *** *CWI, EURANDOM, VU University Amsterdam and Georgia Institute of Technology*

Abstract

In the present paper, we study the evolution of an overloaded cyclic polling model that starts empty. Exploiting a connection with multitype branching processes, we derive fluid asymptotics for the joint queue length process. Under passage to the fluid dynamics, the server switches between the queues infinitely many times in any finite time interval causing frequent oscillatory behavior of the fluid limit in the neighborhood of zero. Moreover, the fluid limit is random. In addition, we suggest a method that establishes finiteness of moments of the busy period in an M/G/1 queue.

Keywords: Cyclic polling; overload; random fluid limit; branching process; multi-stage gated discipline; busy period moment

2010 Mathematics Subject Classification: Primary 60K25; 60F17
Secondary 90B15; 90B22

1. Introduction

This paper is dedicated to stochastic networks called polling models. Broadly speaking, a polling model can be defined as multiple queues served one at a time by a single server. As for further details—service disciplines at the queues, routing of the server, and its walking times from one queue to another—there exist numerous variations motivated by the wide range of applications. The earliest polling study to appear in the literature seems to be the 1957 study by Mack [13], who investigated a problem in the British cotton industry involving a single repairman cyclically patrolling multiple machines, inspecting them for malfunctioning, and repairing them. Over the past few decades, polling techniques have been of extensive use in the areas of computer and communication networks as well as manufacturing and maintenance. Along with that, a vast body of related literature has grown. For overviews of the available results on polling models and their analysis methodologies, we refer the reader to Takagi [17], [18], [19], Boxma [3], Yechiali [25] and Borst [2].

Across the great variety of polling models, there exists the 'classical' one, which was first used in the analysis of time-sharing computer systems in the early 1970s. This model is *cyclic*, i.e. if there are I queues in total, they are visited by the server in the cyclic order $1, 2, \ldots, I, 1, 2, \ldots$. All of the queues are supposed to be infinite-buffer queues, and to each of them there is a Poisson

Received 13 December 2012; revision received 15 April 2013.
* Current address: Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands. Email address: m.remerova@tue.nl
** Postal address: Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, UK.
*** Postal address: CWI, PO Box 94079, 1098 XG Amsterdam, The Netherlands.

stream of customers with independent and identically distributed (i.i.d.) service times. After all visits to a queue, i.i.d. walking, or switchover, times are incurred. All interarrival times, service times and switchover times are mutually independent, and their distributions may vary from queue to queue as well as the service disciplines. Examples of the most common service disciplines are *exhaustive* (the queue is served until it becomes empty), *gated* (in the course of a visit, only those customers get served who are present in the queue when the server arrives at, or *polls*, the queue), and *k-limited* (at most k customers get served per visit). The present paper is also centered around the classical polling model. We assume zero switchover times and allow a wide class of service disciplines that includes both exhaustive and gated policies which we discuss later in more detail.

Amongst desirable properties of any service system, the first is stability. So, naturally, the major part of the polling related literature is focused on the performance of stable models. Foss and Kovalevskii [7] obtained an interesting result of null recurrence over a thick region of parameter space for a two-server modification of polling. MacPhee *et al.* [14], [15] have recently observed the same phenomenon for a hybrid polling/Jackson network, where the service rate and customer rerouting probabilities are randomly updated each time the server switches from one queue to another.

The study of critically loaded polling models was initiated about two decades ago by Coffman *et al.* [4], [5], who proved a so-called averaging principle: in the diffusion heavy-traffic limit, certain functionals of the joint workload process can be expressed via the limit total workload, which was shown to be a reflected Brownian motion and a Bessel process in the case of zero and nonzero switchover times, respectively. In subsequent years, the work has been carried on by Kroese [11], Vatutin and Dyakonova [20], Altman and Kushner [1], van der Mei [21] and others. In particular, heavy-traffic approximations of the steady state and waiting time distributions have been derived.

Although overloaded service systems are an existing reality and it is of importance to control or predict how fast they blow up over time, to the best of the authors' knowledge, for polling models, this problem has not been addressed in the literature so far. The present paper aims to fill in the gap. Moreover, this appears to be a really exciting problem because it reveals the following unusual phenomenon. Our interest is in fluid approximations, i.e. the limit of the scaled joint queue length process

$$(Q_1, \dots, Q_I)\frac{(x^{(n)}\cdot)}{x^{(n)}}$$

along a deterministic sequence $x^{(n)} \to \infty$. Remarkably, in contrast to the many basic queueing systems with deterministic fluid limits, overloaded polling models preserve some randomness under passage to the fluid dynamics. Other examples of simplistic designs combined with random fluid limits are the two-queue two-server models of Foss and Kovalevskii [7] and Kovalevskii *et al.* [10]. We refer the reader to [10] for an insightful discussion of the nature of randomness in fluid limits in general and for an overview of the publications on the topic.

To illustrate the key idea that has led us to the result, consider the simple symmetric model of $I = 2$ queues with exhaustive service, zero switchover times, and empty initial condition (without the last assumption the analysis becomes much simpler). In isolation, the queues are stable and the whole system is overloaded, i.e. $\frac{1}{2} < \lambda/\mu < 1$, where λ and $1/\mu$ are the arrival rate and the mean service time, respectively (in both queues). Denote the supposedly existing limit queue length process by $(\overline{Q}_1, \overline{Q}_2)(\cdot)$. Note that, given the limit size of the queue in service at any nonzero time instant, the entire trajectories of both queues can be restored by the strong law of large numbers (SLLN). Indeed, the limit total population $(\overline{Q}_1 + \overline{Q}_2)(\cdot)$ grows at

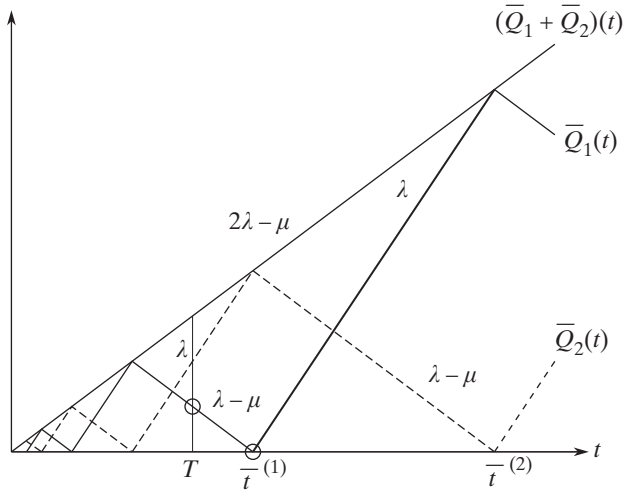

FIGURE 1: Fluid limit of a symmetric two-queue model with exhaustive service.

rate $2\lambda - \mu$. Because of the symmetry, at any fixed time $T > 0$, the queues might (in the limit) be in service with equal probabilities, let it be queue 1. Then in Figure 1 the limit queues 1 and 2 follow the solid and dashed trajectories, respectively. Starting from time T, the limit queue 1 gets cleared up at rate $\lambda - \mu$ until it becomes empty, say, at time $\overline{t}^{(1)}$. After $\overline{t}^{(1)}$, when the limit total population $(2\lambda - \mu)\overline{t}^{(1)}$ comes from queue 2 alone, queue 2 gets cleared up at rate $\lambda - \mu$ until it becomes empty at time $\overline{t}^{(2)}$, while queue 1 grows at the arrival rate λ. Moving forward and backward in this way, we can continue the two trajectories onto $[T, \infty)$ and $(0, T]$, respectively, and see that they oscillate at an infinite rate when approaching zero. Now, the same algorithm applies if $t^{(1)}$ is known, which is the first switching instant after T, and the following crucial observation makes it possible to find the distribution of $t^{(1)}$ (so the randomness of $t^{(1)}$ makes the fluid limit random). Let customer 2 be a descendant of customer 1 if customer 2 arrives at the system while customer 1 is receiving service, or customer 2 is a descendant of a descendant of customer 1. Then the size of the nonempty queue at switching instant forms a branching process.

The idea of representing arriving customers as descendants of the customer in service, has appeared in the work of Foss [6] in the studies of an extension of Klimov's μc-rule, and then in the study by Resing [16], who introduced a wide class of service disciplines that, for the classical polling model (more general periodic server routing is also allowed), guarantee the joint queue length at the successive polling instants of a fixed queue to form a multitype branching process (MTBP). This embedded MTBP is the cornerstone of the analysis that we carry out in this paper.

We now describe the class of service disciplines that we allow in this paper. It is a subclass of the MTBP policies, and we call them *multigated* meaning that each visit to each queue consists of a number of consecutive gated service phases. The upper bound on the number of phases, called the *gating index*, comes from the input data (together with the interarrival and service times). Gating indices for different visits to the same queue are i.i.d. random variables whose distribution may vary from queue to queue, and gating indices for different queues are mutually independent. Gating indices equal to 1 and ∞ correspond to exhaustive and the classical gated service, respectively. Multigated policies with deterministic gating indices were studied (and,

in fact, introduced) recently by van Wijk *et al.* [24] with the purpose of balancing fairness and efficiency of polling models. Van der Mei and co-authors [22], [23] consider multi-stage gated policies, but those are different to those in [24] and here.

Throughout the paper we consider the case of zero switchover times. The case of nonzero switchover times can be treated with similar methods.

As for the proofs, multiple asymmetric queues with nonexhaustive service create more work compared with the simple two-queue example discussed above. Knowing the limit total population is of little use now since it only reduces the dimension of the problem by one. We show that, in our general situation, the fluid limit queue length trajectory $(\overline{Q}_1, \dots, \overline{Q}_I)(\cdot)$ is determined by $2I$ random parameters: the earliest polling instants $\bar{t}_1, \dots, \bar{t}_I$ that, in the limit, follow a fixed time instant, and the limit sizes $\overline{Q}_1(\bar{t}_1), \dots, \overline{Q}_I(\bar{t}_I)$ of the corresponding polled queues. The overload assumption and multigated policies provide the framework of supercritical MTBPs, and we can apply the Kesten–Stigum theorem [9], [12] (the classical result on asymptotics of supercritical MTBPs) to find the distribution of, for example, $(\overline{Q}_1, \dots, \overline{Q}_I)(\bar{t}_1)$. Then suitable SLLNs imply that the other parameters $\bar{t}_1, \dots, \bar{t}_I, \overline{Q}_2(\bar{t}_2), \dots, \overline{Q}_I(\bar{t}_I)$ can be expressed either via the Kesten–Stigum limit $(\overline{Q}_1, \dots, \overline{Q}_I)(\bar{t}_1)$ or via each other. Note also that the Kesten–Stigum theorem requires certain moments of the offspring distribution to be finite. The visit at a queue is the longest when service is exhaustive, implying more customers in the other queues at the end of the cycle. So attempts to satisfy the moment conditions of the Kesten–Stigum theorem boil down to proving finiteness of the corresponding moment for the busy period of an M/G/1 queue, which is an interesting and novel result by itself. In addition, we obtain an estimate for this moment, and our approach is valid for a wide class of regularly varying convex functions, in particular power and logarithmic functions.

The rest of the paper is organized as follows. Section 2 describes the cyclic polling model and the class of service disciplines. Section 3 explains the connection between the model and MTBPs, gives some preliminaries from the theory of MTBPs, and derives characteristics of the embedded MTBP. In Section 4, we state our main result, the fluid limit theorem, and discuss the optimal representation of the fluid limit from the computational point of view (Remark 5). Section 5 proves the results of Section 3; see the proof of Lemma 3 for estimates on the moments of the busy period of an M/G/1 queue. Section 6 proves the fluid limit theorem. Proofs of some auxiliary statements are given in Appendix A.

1.1. Notation

With $x := y$ we mean that x is defined as equal to y.

The standard sets are: positive integers $\mathbb{N} := \{1, 2, \dots\}$; nonnegative integers $\mathbb{Z}_+ := \mathbb{N} \cup \{0\}$; integers $\mathbb{Z} := \{0, \pm 1, \pm 2, \dots\}$.

All vectors are I-dimensional row vectors, $\cdot^\top$ denotes the operation of transposition. All vectors are typeset in bold italic. The vector with all coordinates equal to 0 is denoted by $\mathbf{0}$, with all coordinates equal to 1 by $\mathbf{1}$, and with coordinate i equal to 1 and the other coordinates equal to 0 by $\boldsymbol{e}_i$. The following operations are defined on vectors $\boldsymbol{x} = (x_1, \dots, x_I)$, $\boldsymbol{y} = (y_1, \dots, y_I)$:

- partial order, $\boldsymbol{x} \le \boldsymbol{y}$ if $x_i \le y_i$ for all i;
- L_1-norm $|\boldsymbol{x}| = \sum_{i=1}^I |x_i|$;
- coordinate-wise product $\boldsymbol{x} \times \boldsymbol{y} = (x_1 y_1, \dots, x_I y_I)$;
- power, if all $x_i > 0$, then $\boldsymbol{x}^{\boldsymbol{y}} = \prod_{i=1}^I x_i^{y_i}$;

- binomial coefficient, if $\boldsymbol{y} \le \boldsymbol{x}$, $\binom{\boldsymbol{x}}{\boldsymbol{y}} = \prod_{i=1}^{I} \binom{x_i}{y_i} = \prod_{i=1}^{I} x_i!/y_i!(x_i - y_i)!$.

For a real x, let $\lfloor x \rfloor$ be its maximum integer lower bound, $\lceil x \rceil$ its minimum integer upper bound, and put $\{x\} = x - \lfloor x \rfloor$.

If a superscript is in parentheses, then it is an upper index, otherwise a power.

If random objects X and Y are equal in distribution, we write $X \stackrel{\mathrm{D}}{=} Y$ and say that X is a copy of Y.

2. Model description

This section contains a detailed description of the cyclic polling model and the class of service disciplines that we allow for this model. It also specifies the stochastic assumptions. All stochastic primitives introduced throughout the paper are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$.

2.1. Cyclic polling

Consider a system that consists of multiple infinite-buffer queues labeled by $i = 1, \ldots, I$, where I is finite, and a single server. There are external arrivals of customers to the queues that line up in the corresponding buffers in the order of arrival. The server idles if and only if the entire system is empty. While the system is nonempty, the server works at unit rate serving one queue at a time and switching from one queue to another in the cyclic order: after a period of serving queue i, called a *visit to queue* i, a visit to queue $(i + 1) \bmod I$ follows. Note that, while the system is nonempty, empty queues get visited as well in the sense that, once the server arrives at (or *polls*) an empty queue, say at time t, it has to leave immediately, and the visit in this case is defined to be the empty interval $[t, t)$. Now suppose that, at a particular time instant, the system empties upon completion of a nonempty visit to queue i. For mathematical convenience, we assume that such an instant is followed by a single (empty) visit to each of the empty queues $i + 1, \ldots, I$. Then the server idles until the first arrival into the empty system. If that arrival is to queue i, a single (empty) visit to each of the empty queues $1, \ldots, i - 1$ precedes the visit to queue i. In the course of a visit, a number of customers at the head of the queue get served in the order of arrival and depart. The service disciplines at the queues specify how many customers should get served per visit; we now proceed with their description.

2.2. Multigated service

By multigated service in a queue we mean that each visit to that queue consists of a number of consecutive gated service phases. More formally, we say that *the server gates a queue* at a particular time instant meaning that the queue is in service at the moment, and all of the customers found in the queue at the moment are guaranteed to receive service during the current visit. Customers gated together are served in the order of arrival. For each visit, its *gating index* is defined: it is the upper bound on the number of times the server is supposed to gate the queue in the course of the visit. The gating indices for different queues and for different visits of the same queue might be different. The first time during a visit when the server gates the queue is upon polling the queue. The other gating instants are defined by induction: as soon as the customers found in the queue the last time it was gated have been served, the queue is gated again provided that the total number of gating procedures is not going to exceed the gating index. If the queue is empty upon gating, the server switches to the next queue, and thus the actual number of gating procedures performed during a visit might differ from the gating index for that visit. Now we define a generic multigated service discipline.

Definition 1. Let a random variable X take values in $\mathbb{Z}_+ \cup \{\infty\}$. The service discipline at a particular queue is called *X-gated* if the gating indices for different visits of this queue are i.i.d. copies of X. If a gating index equals 0, the server should leave immediately after polling the queue. The values 1 and ∞ of a gating index correspond to conventional gated and exhaustive service, respectively.

Remark 1. Multigated service disciplines guarantee the population of the polling system, at polling instants of a fixed queue, to be an MTBP, laying the foundation for the analysis that we carry out in this paper. We discuss this connection with MTBPs in detail in the next section.

2.3. Stochastic assumptions

We consider the cyclic polling system described above to evolve in the continuous-time horizon $t \in [0, \infty)$. At $t = 0$, the system is empty. Arrivals of customers to queue i form a Poisson process $E_i(\cdot)$ of rate λ_i. Hence, we introduce the vector of arrival rates

$$\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_I).$$

Service times of queue i customers are drawn from a sequence $\{B_i^{(n)}\}_{n\in\mathbb{N}}$ of i.i.d. copies of a positive random variable B_i with a finite mean value $1/\mu_i$. Gating indices for queue i are drawn from a sequence $\{X_i^{(n)}\}_{n\in\mathbb{N}}$ of i.i.d. copies of a random variable X_i taking values in $\mathbb{Z}_+ \cup \{\infty\}$. The random elements $E_i(\cdot)$, $\{B_i^{(n)}\}_{n\in\mathbb{N}}$, and $\{X_i^{(n)}\}_{n\in\mathbb{N}}$, $i = 1, \ldots, I$, are mutually independent. In addition, we impose the following conditions on the load intensities and service times.

Assumption 1. *For all i, $\lambda_i/\mu_i < 1$, and $\sum_{i=1}^I \lambda_i/\mu_i > 1$.*

Assumption 2. *For all i, $\mathbb{E}B_i \log B_i < \infty$.*

We study the system behavior in terms of its queue length process

$$\boldsymbol{Q}(\cdot) = (Q_1, \ldots, Q_I)(\cdot),$$

where $Q_i(t)$ is the number of customers in queue i at time t.

3. Connection with MTBPs

This section is devoted to a MTBP embedded in the queue length process $\boldsymbol{Q}(\cdot)$ and enabling its further analysis.

To start with, we divide the time horizon into pairwise-disjoint finite intervals in such a way that each interval includes a single (possibly, empty) visit of the server to each of the queues starting from the first one. Let

$$[0, \infty) = \bigcup_{n\in\mathbb{Z}_+} [t^{(n)}, t^{(n+1)}),$$

$$[t^{(n)}, t^{(n+1)}) = [t^{(n)}, t_1^{(n)}) \bigcup_{i=1}^{I} [t_i^{(n)}, t_{i+1}^{(n)}),$$

where:

- $t^{(0)} = 0$ and $t^{(n)} \le t_1^{(n)} \le \cdots \le t_{I+1}^{(n)} = t^{(n+1)}$;
- if the system is empty at $t^{(n)}$, then the interval $[t^{(n)}, t_1^{(n)})$ is the period of waiting until the first arrival, otherwise $t^{(n)} = t_1^{(n)}$;

- the interval $[t_i^{(n)}, t_{i+1}^{(n)})$ is the visit to queue i following $t^{(n)}$, with $t_i^{(n)} = t_{i+1}^{(n)}$ if the queue is empty.

The interval $[t^{(n)}, t^{(n+1)})$ is called *session n*. The interval $[t_i^{(n)}, t_{i+1}^{(n)})$ is called *visit n to queue i*, and the gating index for this visit is $X_i^{(n)}$.

For multigated service disciplines that we consider in this paper, the following holds.

Property 1. *For all $i = 1, \dots, I$, the customers found in queue i at a polling instant get replaced during the course of the visit by i.i.d. copies of a random vector $\check{\boldsymbol{L}}_i = (\check{L}_{i,1}, \dots, \check{L}_{i,I})$ that has the distribution of $\boldsymbol{Q}(t_{i+1}^{(n)})$ given that $\boldsymbol{Q}(t_i^{(n)}) = \boldsymbol{e}_i$ (this distribution does not depend on n).*

By Resing [16], Property 1 implies that the sequence

$$\{\boldsymbol{Q}(t^{(n)})\}_{n \in \mathbb{Z}_+}$$

forms an MTBP with immigration in state $\mathbf{0}$. In the rest of the section, we introduce a number of objects associated with this MTBP and discuss some of its properties.

The random vector $\check{L}_i$, mentioned in Property 1, we call the *visit offspring of a queue i customer*. Define also the *visit duration at queue i* to be a random variable V_i equal in distribution to $t_{i+1}^{(n)} - t_i^{(n)}$ given that $Q_i(t_i^{(n)}) = 1$, and the *session offspring of a queue i customer* to be a random vector $\boldsymbol{L}_i = (L_{i,1}, \dots, L_{i,I})$ that has the distribution of $\boldsymbol{Q}(t^{(n+1)})$ given that $\boldsymbol{Q}(t^{(n)}) = \boldsymbol{e}_i$. Then the immigration distribution is given by

$$G(\boldsymbol{k}) := \mathbb{P}\{\boldsymbol{Q}(t^{(n+1)}) = \boldsymbol{k} \mid \boldsymbol{Q}(t^{(n)}) = \mathbf{0}\} = \frac{\sum_{i=1}^I \lambda_i \mathbb{P}\{\boldsymbol{L}_i = \boldsymbol{k}\}}{\sum_{i=1}^I \lambda_i}, \qquad \boldsymbol{k} \in \mathbb{Z}_+^I.$$

The following lemma computes the mean values

$$\gamma_i := \mathbb{E}V_i, \qquad \check{\boldsymbol{m}}_i = (\check{m}_{i,1}, \dots, \check{m}_{i,I}) := \mathbb{E}\check{\boldsymbol{L}}_i, \qquad \boldsymbol{m}_i = (m_{i,1}, \dots, m_{i,I}) := \mathbb{E}\boldsymbol{L}_i.$$

Lemma 1. *For all i,*

$$\check{m}_{i,i} = \mathbb{E}\left(\frac{\lambda_i}{\mu_i}\right)^{X_i} \quad \text{and} \quad \gamma_i = \frac{1 - \check{m}_{i,i}}{\mu_i - \lambda_i},$$

and, for $i \neq j$,

$$\check{m}_{i,j} = \lambda_j \gamma_i.$$

For the $m_{i,j}$, there is a recursive formula:

$$m_{I,j} = \check{m}_{I,j} \quad \text{for all } j,$$

and, for $i = 1, \dots, I-1$, $\boldsymbol{m}_i$ is computed via $\boldsymbol{m}_{i+1}$,

$$m_{i,j} = \check{m}_{i,j} 1(i \geq j) + \sum_{k=i+1}^{I} \check{m}_{i,k} m_{k,j} \quad \text{for all } j.$$

The proof follows in Section 5.1.

By the Perron–Frobenius theorem (see e.g. [8, Theorem 5.1]), the *mean session offspring matrix* $M := \{m_{i,j}\}_{i,j=1}^I$ has a positive eigenvalue ρ that is greater in absolute value than any other eigenvalue of M. The eigenspace associated to ρ is one-dimensional and parallel to

a vector with all coordinates positive. Consequently, there exist vectors $\boldsymbol{u} = (u_1, \ldots, u_I)$ and $\boldsymbol{v} = (v_1, \ldots, v_I)$ with all coordinates positive such that

$$M\boldsymbol{u}^\top = \rho\boldsymbol{u}^\top, \qquad \boldsymbol{v}M = \rho\boldsymbol{v} \quad \text{and} \quad \boldsymbol{v}\boldsymbol{u}^\top = 1.$$

Now introduce an auxiliary MTBP $\{\boldsymbol{Z}^{(n)}\}_{n\in\mathbb{Z}_+}$ with no immigration and such that, given $\boldsymbol{Z}^{(n)} = \boldsymbol{e}_i$, the next generation $\boldsymbol{Z}^{(n+1)}$ is equal in distribution to $\boldsymbol{L}_i$. Denote by q_i the *extinction probability* for the process $\{\boldsymbol{Z}^{(n)}\}_{n\in\mathbb{Z}_+}$ given that $\boldsymbol{Z}^{(0)} = \boldsymbol{e}_i$, and introduce the vector of extinction probabilities

$$\boldsymbol{q} := (q_1, \ldots, q_I).$$

Then the probability for the process $\{\boldsymbol{Q}(t^{(n)})\}_{n\in\mathbb{Z}_+}$ to return to $\mathbf{0}$ is given by

$$q_G := \sum_{\boldsymbol{k}\in\mathbb{Z}_+^I} G(\boldsymbol{k})\boldsymbol{q}^{\boldsymbol{k}}.$$

Remark 2. Since all time instants t such that $\boldsymbol{Q}(t) = \mathbf{0}$ are contained among the $t^{(n)}$, the probability for the process $\boldsymbol{Q}(\cdot)$ to return to $\mathbf{0}$ also equals q_G.

By Assumption 1, the MTBPs $\{\boldsymbol{Q}(t^{(n)})\}_{n\in\mathbb{Z}_+}$ and $\{\boldsymbol{Z}^{(n)}\}_{n\in\mathbb{Z}_+}$ are supercritical (concerning this, the proof of Lemma 2 is postponed to Appendix A).

Lemma 2. *For the Perron–Frobenius eigenvalue ρ and the extinction probabilities q_i, we have $\rho > 1$ and $q_i < 1$ for all i. By the latter, $q_G < 1$ too.*

Assumption 2 guarantees the finiteness of the corresponding moments for the offspring distribution of the MTBPs $\{\boldsymbol{Q}(t^{(n)})\}_{n\in\mathbb{Z}_+}$ and $\{\boldsymbol{Z}^{(n)}\}_{n\in\mathbb{Z}_+}$ (see Section 5.2 for the proof of Lemma 3).

Lemma 3. *For all i and j, $\mathbb{E}L_{i,j}\log L_{i,j} < \infty$, where $0\log 0 := 0$ by convention.*

Finally, we utilize the Kesten–Stigum theorem for supercritical MTBPs (see, e.g. [9] and [12]). It is our starting point when proving the convergence results of the next section. By that theorem and Lemmas 2 and 3, the auxiliary process $\{\boldsymbol{Z}^{(n)}\}_{n\in\mathbb{Z}_+}$ has the following asymptotics.

Proposition 1. *Given $\boldsymbol{Z}^{(0)} = \boldsymbol{e}_i$,*

$$\frac{\boldsymbol{Z}^{(n)}}{\rho^n} \to \zeta_i\boldsymbol{v} \quad \textit{almost surely (a.s.) as } n \to \infty,$$

where the distribution of the random variable ζ_i has a jump of magnitude $q_i < 1$ at 0 and a continuous density function on $(0, \infty)$, and $\mathbb{E}\zeta_i = u_i$.

4. Fluid limit

In this section, we present out main result which concerns the behavior of the model under study on a large time scale.

For each $n \in \mathbb{Z}_+$, we introduce the scaled queue length process

$$\overline{\boldsymbol{Q}}^{(n)}(t) := \frac{\boldsymbol{Q}(\rho^n t)}{\rho^n}, \qquad t \in [0, \infty). \tag{1}$$

We are interested in the almost sure limit of the processes (1) as $n \to \infty$, which we call the *fluid limit* of the model. It appears that, in order to precisely describe the fluid limit, the information provided by Theorem 1 below is sufficient.

For $n \in \mathbb{Z}$, let

$$\eta_n := \begin{cases} \min\{k\colon t^{(k)} \geq \rho^n\} & \text{if } n \geq 0, \\ 0 & \text{if } n < 0. \end{cases}$$

Theorem 1. *There exist constants $\overline{b}_i \in (0,\infty)$ and $\overline{\boldsymbol{a}}_i = (\overline{a}_{i,1},\dots,\overline{a}_{i,I}) \in [0,\infty)^I$, $i = 1,\dots,I+1$, and a random variable ξ with values in $[1,\rho)$ such that, for all $k \in \mathbb{Z}_+$ and i,*

$$\frac{t_i^{(\eta_n+k)}}{\rho^n} \to \rho^k \overline{b}_i \xi \quad \text{and} \quad \frac{\boldsymbol{Q}\left(t_i^{(\eta_n+k)}\right)}{\rho^n} \to \xi \rho^k \overline{\boldsymbol{a}}_i \quad \text{a.s. as } n \to \infty. \tag{2}$$

The $\overline{b}_i$ and $\overline{\boldsymbol{a}}_i$ are given by

$$\overline{b}_1 = 1, \qquad \overline{b}_{i+1} = \overline{t}_i + \left(\frac{v_i}{\alpha} + \lambda_i(\overline{b}_i - \overline{b}_1)\right)\gamma_i, \qquad i = 1,\dots,I, \tag{3}$$

and

$$\overline{\boldsymbol{a}}_1 = \frac{v}{\alpha}, \qquad \overline{\boldsymbol{a}}_{i+1} = \overline{\boldsymbol{a}}_i + (\overline{b}_{i+1} - \overline{b}_i)\boldsymbol{\lambda} - (\overline{b}_{i+1} - \overline{b}_i)\mu_i \boldsymbol{e}_i, \qquad i = 1,\dots,I, \tag{4}$$

where

$$\alpha = \frac{\sum_{i=1}^I v_i/\mu_i}{\sum_{i=1}^I \lambda_i/\mu_i - 1}.$$

For $x \in [1,\rho)$, the distribution of ξ is given by

$$\mathbb{P}\{\xi \geq x\} = \frac{1}{1-q_G} \sum_{\substack{\boldsymbol{k}\in\mathbb{Z}_+^I, \\ |\boldsymbol{k}|\geq 1}} G(\boldsymbol{k}) \sum_{\substack{\boldsymbol{l}\leq\boldsymbol{k}, \\ |\boldsymbol{l}|\geq 1}} \binom{\boldsymbol{k}}{\boldsymbol{l}} (\boldsymbol{1}-\boldsymbol{q})^{\boldsymbol{l}} \boldsymbol{q}^{\boldsymbol{k}-\boldsymbol{l}} \times \mathbb{P}\left\{\left\{\log_\rho\left(\alpha \sum_{i=1}^I \sum_{j=1}^{l_i} \xi_i^{(j)}\right)\right\} \geq \log_\rho x\right\},$$

where $\xi_i^{(j)}$, $j \in \mathbb{N}$, are i.i.d. random variables with the distribution of ζ_i given that $\zeta_i > 0$, and the sequences $\{\xi_i^{(j)}\}_{j\in\mathbb{N}}$, $i = 1,\dots,I$, are mutually independent.

The proof of Theorem 1 combines the Kesten–Stigum theorem with various dynamic equations and laws of large numbers, see Section 6.

Remark 3. Since $t_{I+1}^{(n)} = t_1^{(n)}$, we also have

$$\overline{b}_{I+1} = \rho \overline{b}_1 \quad \text{and} \quad \overline{\boldsymbol{a}}_{I+1} = \rho \overline{\boldsymbol{a}}_1.$$

Remark 4. There is an alternative way to compute the $\overline{a}_i$:

$$\overline{\boldsymbol{a}}_1 = v, \qquad \overline{\boldsymbol{a}}_{i+1} = \overline{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}_{i,i}\boldsymbol{e}_i + \overline{a}_{i,i}\check{\boldsymbol{m}}_i, \qquad i = 1,\dots,I,$$

which implies that $\overline{a}_{i,j} > 0$ if $|i-j| \neq 1$ and $\overline{a}_{i,i+1} = 0$ if and only if the service discipline at queue i is exhaustive. See Lemma 7 and Remark 6 in Section 6.2.

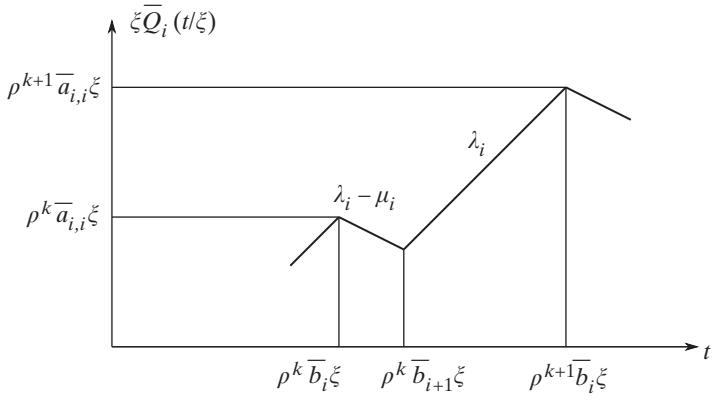

FIGURE 2: Fluid limit of queue i.

Based on the results of Theorem 1, Theorem 2 below derives the fluid limit equations from the suitable dynamic equations, see Section 6 for the proof.

Theorem 2. *There exists a deterministic function*

$$\overline{\boldsymbol{Q}}(\cdot) = (\overline{Q}_1, \ldots, \overline{Q}_I)(\cdot)\colon [0,\infty) \to [0,\infty)^I$$

such that, a.s. as $n \to \infty$,

$$\overline{\boldsymbol{Q}}^{(n)}(\cdot) \to \xi\,\overline{\boldsymbol{Q}}\left(\frac{\cdot}{\xi}\right)$$

uniformly on compact sets, where the random variable ξ *is defined in Theorem 1.*

The function $\overline{\boldsymbol{Q}}(\cdot)$ *is continuous and piecewise linear and given by*

$$\overline{\boldsymbol{Q}}(t) = \begin{cases} 0 & \text{if } t = \mathbf{0}, \\ \rho^k\overline{\boldsymbol{a}}_i + (t-\rho^k\overline{b}_i)\boldsymbol{\lambda} - (t-\rho^k\overline{b}_i)\mu_i\boldsymbol{e}_i & \text{if } t \in [\rho^k\overline{b}_i, \rho^k\overline{b}_{I+1}), \\ & \quad i = 1,\ldots,I,\ k \in \mathbb{Z}, \end{cases} \tag{5}$$

or, equivalently, by

$$\text{for all } i, \quad \overline{Q}_i(t) = \begin{cases} 0 & \text{if } t = 0, \\ \rho^k\overline{a}_{i,i} + (\lambda_i-\mu_i)(t-\rho^k\overline{b}_i) & \text{if } t \in [\rho^k\overline{b}_i, \rho^k\overline{b}_{i+1}),\ k \in \mathbb{Z}, \\ \rho^{k+1}\overline{a}_{i,i} - \lambda_i(\rho^{k+1}\overline{b}_i - t) & \text{if } t \in [\rho^k\overline{b}_{i+1}, \rho^{k+1}\overline{b}_i),\ k \in \mathbb{Z}. \end{cases} \tag{6}$$

Remark 5. By (6), the whole process $\overline{\boldsymbol{Q}}(\cdot)$ is defined by the constants $\overline{b}_i$ and $\overline{a}_{i,i}$. The fastest way to compute the $\overline{b}_i$ and $\overline{a}_{i,i}$ is using the simultaneous recursion

$$\overline{b}_1 = 1, \qquad \overline{a}_{i,i} = \frac{v_i}{\alpha} + \lambda_i(\overline{b}_i - \overline{b}_1), \qquad \overline{b}_{i+1} = \overline{b}_i + \overline{a}_{i,i}\gamma_i, \qquad i = 1,\ldots,I.$$

See the last part of the proof of Lemma 7 (namely, (32) and (33)) and Remark 6 in Section 6.2.

Finally, Figure 2 depicts a trajectory of the limiting process $\xi\overline{Q}(\cdot/\xi)$.

5. Proofs for Section 3

In this section, we prove the properties of the offspring distribution of the embedded MTBP $\{\boldsymbol{Q}(t^{(n)})\}_{n\in\mathbb{Z}_+}$.

5.1. Proof of Lemma 1

First we compute the γ_i. For $k \in \mathbb{Z}_+ \cup \{\infty\}$, let a random variable $V_i(k)$ be the visit duration at queue i given that the service discipline at queue i is k-gated. Recall that the gating index ∞ corresponds to exhaustive service, and hence

$$\mathbb{E}V_i(\infty) = \frac{1}{\mu_i - \lambda_i}.$$

Now note that

$$V_i(0) = 0 \quad \text{and} \quad V_i(k+1) \stackrel{\mathrm{D}}{=} B_i + \sum_{l=1}^{E_i(B_i)} V_i^{(l)}(k), \qquad k \in \mathbb{Z}_+. \tag{7}$$

where the random elements B_i, $E_i(\cdot)$ and $\{V_i^{(l)}(k)\}_{l\in\mathbb{N}}$ are mutually independent, and $V_i^{(l)}(k)$, $l \in \mathbb{N}$, are i.i.d. copies of $V_i(k)$. Then, for $k \in \mathbb{Z}_+$,

$$\begin{aligned}
\mathbb{E}V_i(k+1) &= \frac{1}{\mu_i} + \frac{\lambda_i}{\mu_i}\mathbb{E}V_i(k) = \frac{1}{\mu_i}\left(1 + \frac{\lambda_i}{\mu_i}\right) + \left(\frac{\lambda_i}{\mu_i}\right)^2 \mathbb{E}V_i(k-1) \\
&= \cdots \\
&= \frac{1}{\mu_i}\left(1 + \frac{\lambda_i}{\mu_i} + \cdots + \left(\frac{\lambda_i}{\mu_i}\right)^k\right) + \left(\frac{\lambda_i}{\mu_i}\right)^{k+1} \mathbb{E}V_i(0) \\
&= \frac{1}{\mu_i}\frac{1 - (\lambda_i/\mu_i)^{k+1}}{1 - \lambda_i/\mu_i},
\end{aligned} \tag{8}$$

and

$$\gamma_i = \sum_{k\in\mathbb{Z}_+\cup\{\infty\}} \mathbb{P}\{X_i = k\}\mathbb{E}V_i(k) = \frac{1 - \mathbb{E}(\lambda_i/\mu_i)^{X_i}}{\mu_i - \lambda_i}.$$

In a similar way, we compute the $\check{m}_{i,i}$. For $k \in \mathbb{Z}_+ \cup \{\infty\}$, let a random variable $\check{L}_{i,i}(k)$ be the queue i visit offspring of a queue i customer given that the service discipline at queue i is k-gated. Since

$$\check{L}_{i,i}(\infty) = 0, \qquad \check{L}_{i,i}(0) = 1, \quad \text{and} \quad \check{L}_{i,i}(k+1) \stackrel{\mathrm{D}}{=} \sum_{l=1}^{E_i(B_i)} \check{L}_{i,i}^{(l)}(k), \qquad k \in \mathbb{Z}_+,$$

where the random elements B_i, $E_i(\cdot)$ and $\{\check{L}_{i,i}^{(l)}(k)\}_{l\in\mathbb{N}}$ are mutually independent, and $\check{L}_{i,i}^{(l)}(k)$, $l \in \mathbb{N}$, are i.i.d. copies of $\check{L}_{i,i}(k)$, we have

$$\mathbb{E}\check{L}_{i,i}(k+1) = \frac{\lambda_i}{\mu_i}\mathbb{E}\check{L}_{i,i}(k) = \cdots = \left(\frac{\lambda_i}{\mu_i}\right)^{k+1}, \qquad k \in \mathbb{Z}_+,$$

$$\check{m}_{i,i} = \mathbb{E}\left(\frac{\lambda_i}{\mu_i}\right)^{X_i}.$$

The formulas for the $\check{m}_{i,j}$, $i \neq j$, and the $m_{i,j}$ follow, respectively, by the representations

$$\check{L}_{i,j} \stackrel{\mathrm{D}}{=} E_j(V_i), \qquad i \neq j, \tag{9}$$

where V_i and $E_j(\cdot)$ are independent, and

$$L_{i,j} \stackrel{\mathrm{D}}{=} \begin{cases} \check{L}_{i,j} + \sum\limits_{l=1}^{\check{L}_{i,i+1}} L^{(l)}_{i+1,j} + \cdots + \sum\limits_{l=1}^{\check{L}_{i,I}} L^{(l)}_{I,j}, & i \geq j, \\ \sum\limits_{l=1}^{\check{L}_{i,i+1}} L^{(l)}_{i+1,j} + \cdots + \sum\limits_{l=1}^{\check{L}_{i,I}} L^{(l)}_{I,j}, & i < j, \end{cases} \tag{10}$$

where $L^{(l)}_{i,j}$, $l \in \mathbb{N}$, are i.i.d. copies of $L_{i,j}$, and the sequences $\{L^{(l)}_{i,j}\}_{l\in\mathbb{N}}$, $i, j = 1, \ldots, I$, are mutually independent and do not depend on the vectors $\check{L}_i$, $i = 1, \ldots, I$.

5.2. Proof of Lemma 3

The cornerstone of this proof is finiteness of the corresponding moments for the busy periods of the queues in isolation, which we check with the help of the auxiliary Lemmas 4 and 5 that follow below together with their proofs.

Lemma 4. *Suppose that a function $f(\cdot)\colon [0,\infty) \to [0,\infty)$ is bounded in a finite interval $[0, T]$ and nondecreasing in $[T, \infty)$, and that $f(x) \to \infty$ as $x \to \infty$. Suppose also that, for some (and, hence, for all) $c > 1$,*

$$\limsup_{x\to\infty} \frac{f(cx)}{f(x)} < \infty. \tag{11}$$

Consider an i.i.d. sequence $\{Y^{(n)}\}_{n\in\mathbb{N}}$ of nonnegative, nondegenerate at 0 random variables, and the renewal process

$$Y(t) = \max\left\{n \in \mathbb{Z}_+\colon \sum_{k=1}^{n} Y^{(k)} \leq t\right\}, \qquad t \in [0,\infty).$$

Let τ be a nonnegative random variable which may depend on the sequence $\{Y_n\}_{n\in\mathbb{N}}$. Assume that $\mathbb{E} f(\tau) < \infty$. Then $\mathbb{E} f(Y(\tau))$ is finite too.

Proof. Without loss of generality, we can assume that the function $f(\cdot)$ is nondecreasing in the entire domain $[0,\infty)$ (otherwise, instead of $f(\cdot)$, one can consider $\widetilde{f}(\cdot) = \sup_{0\leq y\leq \cdot} f(y)$), and also that $f(\cdot)$ is right-continuous.

First we show that, if (11) holds for some $c > 1$, then it holds for any $c' > 1$. For $c' = c^k$, $k \in \mathbb{N}$, we have

$$\limsup_{x\to\infty} \frac{f(c^k x)}{f(x)} \leq \limsup_{x\to\infty} \frac{f(c^k x)}{f(c^{k-1}x)} \limsup_{x\to\infty} \frac{f(c^{k-1}x)}{f(c^{k-2}x)} \cdots \limsup_{x\to\infty} \frac{f(cx)}{f(x)} < \infty.$$

Then, for $c' > 1$ other than powers of c, (11) follows by the monotonicity of $f(\cdot)$.

Condition (11) also implies that

$$\lim_{x\to\infty} \frac{\log(f(x))}{x} = 0. \tag{12}$$

Indeed, in (11) take $c = e$, the exponent. Since $M := \limsup_{x\to\infty} f(ex)/f(x) < \infty$, there exists a large enough $T' > 0$ such that $\sup_{x\in[T',\infty)} f(ex)/f(x) \le 2M$. Note that any $x \in [eT', \infty)$ admits a unique representation $x = \mathrm{e}^{k(x)}y(x)$, where $y(x) \in [T', eT')$ and $k(x) \in \mathbb{N}$. Hence, for any $x \in [eT', \infty)$,

$$f(x) = \frac{f(\mathrm{e}^{k(x)}y(x))}{f(\mathrm{e}^{k(x)-1}y(x))}\frac{f(\mathrm{e}^{k(x)-1}y(x))}{f(\mathrm{e}^{k(x)-2}y(x))}\cdots\frac{f(ey(x))}{f(y(x))}f(y(x)) \le (2M)^{k(x)}f(eT')$$

and

$$\frac{\log(f(x))}{x} \le \frac{k(x)\log(2M) + \log(f(eT'))}{T'\mathrm{e}^{k(x)}},$$

implying (12).

Now define the pseudo-inverse function

$$f^{(-1)}(y) := \inf\{x \in [0,\infty)\colon f(x) \ge y\}, \qquad y \in [0,\infty).$$

For any $c' > 0$, we have

$$\begin{aligned}
\mathbb{E}f(Y(\tau)) &\le \sum_{n\in\mathbb{Z}_+}\mathbb{P}\{f(Y(\tau)) \ge n\}\\
&\le \sum_{n\in\mathbb{Z}_+}\mathbb{P}\{Y(\tau) \ge f^{(-1)}(n)\} \le \sum_{n\in\mathbb{Z}_+}\mathbb{P}\left\{\sum_{k=1}^{\lceil f^{(-1)}(n)\rceil} Y^{(k)} \le \tau\right\}\\
&\le \underbrace{\sum_{n\in\mathbb{Z}_+}\mathbb{P}\left\{\sum_{k=1}^{\lceil f^{(-1)}(n)\rceil} Y^{(k)} \le c'\lceil f^{(-1)}(n)\rceil\right\}}_{=:\,\Sigma_1(c')} + \underbrace{\sum_{n\in\mathbb{Z}_+}\mathbb{P}\{c'f^{(-1)}(n) < \tau\}}_{=:\,\Sigma_2(c')}.
\end{aligned}$$

By condition (11), $\mathbb{E}f(\tau/c') < \infty$ for any $c' > 0$ and, hence,

$$\Sigma_2(c') \le \sum_{n\in\mathbb{Z}_+}\mathbb{P}\left\{f\left(\frac{\tau}{c'}\right) \ge n\right\} \le 1 + \mathbb{E}f\left(\frac{\tau}{c'}\right) < \infty.$$

We now pick a c' such that $\Sigma_1(c') < \infty$, and this will finish the proof. By Markov's inequality, $\mathbb{P}\{\sum_{k=1}^n Y^{(k)} \le c'n\} = \mathbb{P}\{\mathrm{e}^{-\sum_{k=1}^n Y^{(k)}} \ge \mathrm{e}^{-c'n}\} \le (\mathrm{e}^{c'}\mathbb{E}\mathrm{e}^{-Y^{(1)}})^n$. Let c' be small enough so that $\widetilde{c} := \mathrm{e}^{c'}\mathbb{E}\mathrm{e}^{-Y^{(1)}} < 1$. Since $\lceil f^{(-1)}(n)\rceil = m$ implies $n \le f(m+1)$, we have

$$\Sigma_1(c') \le \sum_{m\in\mathbb{Z}_+}\mathbb{P}\left\{\sum_{k=1}^m Y^{(k)} \le c'm\right\}f(m+1) \le \frac{1}{\widetilde{c}}\sum_{m\in\mathbb{N}}\widetilde{c}^m f(m).$$

Take an $\varepsilon \in (0, |\log(\widetilde{c})|)$. By (12), there exists a large enough $N \in \mathbb{N}$ such that $f(m) \le \mathrm{e}^{m\varepsilon}$ for $m > N$. Then

$$\Sigma_1(c') \le \frac{1}{\widetilde{c}}\sum_{m=1}^N \widetilde{c}^m f(m) + \frac{1}{\widetilde{c}}\sum_{m=N+1}^\infty (\widetilde{c}\mathrm{e}^{\varepsilon})^m,$$

where $\widetilde{c}\mathrm{e}^{\varepsilon} = \mathrm{e}^{\varepsilon - |\log(\widetilde{c})|} < 1$ by the choice of ε and, hence, $\Sigma_1(c') < \infty$.

Lemma 5. *Consider a sequence* $\{Y^{(n)}\}_{n\in\mathbb{N}}$ *of nonnegative random variables that are identically distributed* (*but not necessarily independent*)*, and also a* $\mathbb{Z}_+$*-valued random variable* η *that does not depend on* $\{Y^{(n)}\}_{n\in\mathbb{N}}$*. If* $f(\cdot)\colon [0,\infty)\to\mathbb{R}$ *is a convex function, then*

$$\mathbb{E}f\left(\sum_{k=1}^{\eta} Y^{(k)}\right) \le \mathbb{E}f(\eta Y^{(1)}).$$

Proof. By the convexity of $f(\cdot)$, for any $n\in\mathbb{Z}_+$,

$$\mathbb{E}f\left(\sum_{k=1}^{n} Y^{(k)}\right) = \mathbb{E}f\left(\sum_{k=1}^{n} \frac{1}{n}(nY^{(k)})\right) \le \sum_{k=1}^{n}\frac{1}{n}\mathbb{E}f(nY^{(k)}) = \mathbb{E}f(nY^{(1)}).$$

Then Lemma 5 follows by the independence between $\{Y^{(n)}\}_{n\in\mathbb{N}}$ and η:

$$\begin{aligned}\mathbb{E}f\left(\sum_{k=1}^{\eta} Y^{(k)}\right) &= \sum_{n\in\mathbb{Z}_+}\mathbb{P}\{\eta=n\}f\left(\sum_{k=1}^{n} Y^{(k)}\right)\\ &\le \sum_{n\in\mathbb{Z}_+}\mathbb{P}\{\eta=n\}\,\mathbb{E}f(nY^{(1)}) = \mathbb{E}f(\eta Y^{(1)}).\end{aligned}$$

Now we proceed with the proof of Lemma 3. It suffices to show that

$$\mathbb{E}f(L_{i,j}) < \infty, \quad \text{for all } i \text{ and } j,$$

where

$$f(x) = \begin{cases} 0, & x\in[0,1],\\ x\log x, & x\in[1,\infty).\end{cases}$$

Note that the function $f(\cdot)$ is convex: in $(1,\infty)$, its derivative $\log(\cdot)+1$ is nondecreasing, and in the other points, it is easy to check the definition of convexity. Also note that

$$f(xy) \le xf(y) + yf(x), \qquad x, y\in[0,\infty). \tag{13}$$

The rest of the proof is divided into three parts. The two key steps are to show that the f-moments of the visit duration V_i and the same type visit offspring $\check{L}_{i,i}$ are finite. Then the finiteness of the f-moments of the session offspring $L_{i,j}$ follows easily.

5.2.1. *Finiteness of* $\mathbb{E}f(V_i)$. It suffices to show that, in the M/G/1 queue with the arrival process $E_i(\cdot)$ and service times $B_i^{(n)}$, $n\in\mathbb{N}$, the f-moment of the busy period is finite. Suppose that at time $t=0$, there is one customer in the queue, and his/her service time $B_i^{(0)}$ is equal in distribution to B_i and is independent from $E_i(\cdot)$ and $\{B_i^{(n)}\}_{n\in\mathbb{N}}$. Let

$$\tau_i = \min\{t\in(0,\infty)\colon \text{the queue is empty at } t\},$$

$$\tau_i(0)=0, \qquad \tau_i(1)=B_i^{(0)}, \qquad \tau_i(k+2)=\tau_i(k+1)+\sum_{n=E_i(\tau_i(k))+1}^{E_i(\tau_i(k+1))} B_i^{(n)}, \qquad k\in\mathbb{Z}_+.$$

Whilst τ_i is a busy period, $\tau_i(k)$ is equal in distribution to the visit duration in queue i of the polling system given that the service discipline in that queue is k-gated, and

$$\tau_i(k)\uparrow\tau_i \quad \text{a.s. as } k\to\infty.$$

Now we show that the moments $\mathbb{E}f(\tau_i(k)), k \in \mathbb{Z}_+$, are bounded. Then the finiteness of $\mathbb{E}f(\tau_i)$ follows by the continuity of $f(\cdot)$ and the dominated convergence theorem.

Mimicking (7) , we have

$$\tau_i(k+1) \stackrel{\mathrm{D}}{=} B_i^{(0)} + \sum_{l=1}^{E_i(B_i^{(0)})} \tau_i(k)^{(l)}, \qquad k \geq 1,$$

where $\tau_i(k)^{(l)}, l \in \mathbb{N}$, are i.i.d. copies of $\tau_i(k)$ that are independent from $B_i^{(0)}$ and $E_i(\cdot)$. Then, by the monotonicity and convexity of $f(\cdot)$, and the auxiliary Lemma 5 combined with (13),

$$\begin{aligned}
\mathbb{E}f(\tau_i(k)) \leq \mathbb{E}f(\tau_i(k+1)) &\leq \frac{1}{2}\mathbb{E}f(2B_i^{(0)}) + \frac{1}{2}\mathbb{E}f\left(2\sum_{l=1}^{E_i(B_i^{(0)})} \tau_i(k)^{(l)}\right) \\
&\leq \frac{1}{2}\mathbb{E}f(2B_i^{(0)}) + \frac{1}{2}\mathbb{E}f(2E_i(B_i^{(0)})\tau_i(k)^{(1)}) \\
&\leq \frac{1}{2}\mathbb{E}f(2B_i^{(0)}) + \frac{\lambda_i}{\mu_i}\mathbb{E}f(\tau_i(k)) + \frac{1}{2}\mathbb{E}\tau_i(k)\mathbb{E}f(2E_i(B_i^{(0)})),
\end{aligned}$$

where $\mathbb{E}f(2E_i(B_i^{(0)})) < \infty$ by the auxiliary Lemma 4, and $\mathbb{E}\tau_i(k) \leq 1/(\mu_i - \lambda_i)$ by (8). Thus, for all $k \geq 2$,

$$\mathbb{E}f(\tau_i(k)) \leq \frac{c}{1 - \lambda_i/\mu_i},$$

where

$$c = \frac{\mathbb{E}f(2B_i^{(0)})}{2} + \frac{\mathbb{E}f(2E_i(B_i^{(0)}))}{2(\mu_i - \lambda_i)} < \infty.$$

5.2.2. *Finiteness of* $\mathbb{E}f(\check{L}_{i,i})$. Note that L_{ii} is bounded stochastically from above by the number of service completions during the busy period of the M/G/1 queue introduced when proving the finiteness of $\mathbb{E}f(V_i)$. The number of service completions during the first busy period τ_i is given by $1 + E_i(\tau_i)$, and the finiteness of $\mathbb{E}f(1 + E_i(\tau_i))$ follows by the auxiliary Lemma 4.

5.2.3. *Finiteness of* $\mathbb{E}f(L_{i,j})$. This part of the proof uses mathematical induction. Now that we have shown the finiteness of the moments $\mathbb{E}f(\check{L}_{i,i})$, (9) and Lemma 4 imply that

$$\mathbb{E}f(\check{L}_{i,j}) < \infty \quad \text{for all } i \text{ and } j. \tag{14}$$

Then we have the basis of induction: $\mathbb{E}f(L_{I,j}) = \mathbb{E}f(\check{L}_{I,j}) < \infty$ for all j. Suppose that $\mathbb{E}f(L_{k,j}) < \infty$ for $k = i+1, \ldots, I$ and all j. Then the induction step (from $i+1$ to i) follows by (10), the convexity of $f(\cdot)$, Lemma 5 combined with (13), and (14).

6. Proofs for Section 4

First we make preparations in Sections 6.1 and 6.2, and then proceed with the proofs of Theorems 1 and 2 in Sections 6.3 and 6.4, respectively.

6.1. Additional notation

In this section we introduce a number of auxiliary random objects that we operate with when proving the almost sure convergence results of the paper.

6.1.1. *Queue length dynamics.* Define the renewal processes

$$B_i(t) := \max\left\{n \in \mathbb{Z}_+ \text{ such that } \sum_{k=1}^{n} B_i^{(k)} \le t\right\}, \qquad t \in [0, \infty),$$

and the processes

$$I_i(t) := \int_0^t 1(\text{queue } i \text{ is in service at time } s)\, \mathrm{d}s, \qquad t \in [0, \infty),$$

that keep track of how much time the server has spent in each of the queues. Then the number of queue i customers that have departed up to time t is given by

$$D_i(t) := B_i(I_i(t)).$$

Most of the almost sure convergence results of the paper we derive from the basic equations

$$Q_i(\cdot) = E_i(\cdot) - D_i(\cdot).$$

The preliminary results of Section 6.2 depend on when the system empties for the last time. The number of indices n such that $\boldsymbol{Q}(t^{(n)}) = 0$ has a geometric distribution with parameter $q_G < 1$ (see Lemma 2). Denote by ν the last such index, i.e.

$$\boldsymbol{Q}(t^{(\nu)}) = \mathbf{0} \quad \text{and} \quad \boldsymbol{Q}(t^{(n)}) \neq \mathbf{0} \quad \text{for all } n > \nu.$$

6.1.2. *Ancestor-descendant relationships between customers.* By the following three rules, we define the binary relation '*is a descendant of*' on the set of customers:

- each customer is a descendant of himself/herself;
- if customer 2 arrives while customer 1 is receiving service (the two customers are allowed to come from different queues), then customer 2 is a descendant of customer 1;
- if customer 2 is a descendant of customer 1, and customer 3 a descendant of customer 2, then customer 3 is a descendant of customer 1.

Now suppose that a customer is in position k in queue i at the beginning of visit n to queue i. Denote by $V_i^{(n,k)}$ the amount of time during the visit that his/her descendants are in service, and by $\check{L}_{i,j}^{(n,k)}$ the number of his/her descendants in queue j at the end of the visit. If a customer is in position k in queue i at the beginning of session n, denote by $L_{i,j}^{(n,k)}$ the number of his/her descendants in queue j at the end of the session. Introduce also the random vectors

$$\check{\boldsymbol{L}}_i^{(n,k)} := (\check{L}_{i,1}^{(n,k)}, \ldots, \check{L}_{i,I}^{(n,k)}) \quad \text{and} \quad \boldsymbol{L}_i^{(n,k)} := (L_{i,1}^{(n,k)}, \ldots, L_{i,I}^{(n,k)}).$$

6.2. Preliminary results

In this section, we characterize the asymptotic behavior of the system at the switching instants $t_i^{(n)}$, laying the basis for Theorem 1 that concerns the bigger scale times $t_i^{(\eta_n)}$.

From the Kesten–Stigum theorem, we derive the following result for the $t_1^{(n)}$.

Lemma 6. *There exists a positive random variable ζ such that*

$$\frac{\boldsymbol{Q}(t^{(n)})}{\rho^n} \to \zeta \boldsymbol{v} \quad \text{and} \quad \frac{\boldsymbol{Q}(t_1^{(n)})}{\rho^n} \to \zeta \boldsymbol{v} \quad \text{a.s. as } n \to \infty.$$

For $x \in (0, \infty)$, the distribution of ζ is given by

$$\mathbb{P}\{\zeta \geq x\} = \frac{1}{1-q_G} \sum_{n \in \mathbb{Z}_+} \mathbb{P}\{\nu = n\} \sum_{\substack{\boldsymbol{k} \in \mathbb{Z}_+^I, \\ |\boldsymbol{k}| \geq 1}} G(\boldsymbol{k}) \sum_{\substack{\boldsymbol{l} \leq \boldsymbol{k}, \\ |\boldsymbol{l}| \geq 1}} \binom{\boldsymbol{k}}{\boldsymbol{l}} (\boldsymbol{1}-\boldsymbol{q})^{\boldsymbol{l}} \boldsymbol{q}^{\boldsymbol{k}-\boldsymbol{l}} \times \mathbb{P}\left\{ \sum_{i=1}^{I} \sum_{j=1}^{l_i} \xi_i^{(j)} \geq \rho^{n+1} x \right\}, \quad (15)$$

where the random variables $\xi_i^{(j)}$ are the same as in Theorem 1.

Proof. Since $t_1^{(n)} = t^{(n)}$ for $n > \nu$, it suffices to find the almost sure limit of $\boldsymbol{Q}(t^{(n)})/\rho^n$.

First we find the asymptotics of the auxiliary MTBP $\{\boldsymbol{Z}^{(n)}\}_{n \in \mathbb{N}}$ (without immigration) under the assumption that $\boldsymbol{Z}^{(0)}$ is distributed according to $\{G(\boldsymbol{k})\}_{\boldsymbol{k} \in \mathbb{Z}_+^I}$ (the immigration distribution for the MTBP $\{\boldsymbol{Q}(t^{(n)})\}_{n \in \mathbb{N}}$).

By Proposition 1, if the distribution of $\boldsymbol{Z}^{(0)}$ is $\{G(\boldsymbol{k})\}_{\boldsymbol{k} \in \mathbb{Z}_+^I}$, we have

$$\frac{\boldsymbol{Z}^{(n)}}{\rho^n} \to \underbrace{\left(\sum_{\boldsymbol{k} \in \mathbb{Z}_+^I} 1(\boldsymbol{Z}^{(0)} = \boldsymbol{k}) \sum_{i=1}^{I} \sum_{j=1}^{k_i} \zeta_i^{(j)} \right)}_{=: \zeta_G} \boldsymbol{v} \quad \text{a.s. as } n \to \infty,$$

where $\zeta_i^{(j)}$, $j \in \mathbb{N}$, are i.i.d. copies of ζ_i, and the sequences $\{\zeta_i^{(j)}\}_{j \in \mathbb{N}}$, $i = 1, \ldots, I$, are mutually independent and also independent from $\boldsymbol{Z}^{(0)}$.

With $x \in (0, \infty)$, the distribution of ζ_G is given by

$$\mathbb{P}\{\zeta_G = 0\} = q_G,$$

$$\mathbb{P}\{\zeta_G \geq x\} = \sum_{\substack{\boldsymbol{k} \in \mathbb{Z}_+^I, \\ |\boldsymbol{k}| \geq 1}} G(\boldsymbol{k}) \sum_{\substack{\boldsymbol{l} \leq \boldsymbol{k}, \\ |\boldsymbol{l}| \geq 1}} \binom{\boldsymbol{k}}{\boldsymbol{l}} p(\boldsymbol{l}) \underbrace{\prod_{i=1}^{I} \mathbb{P}\{\zeta_i > 0\}^{l_i} \mathbb{P}\{\zeta_i = 0\}^{k_i - l_i}}_{= (\boldsymbol{1}-\boldsymbol{q})^{\boldsymbol{k}-\boldsymbol{l}} \boldsymbol{q}^{\boldsymbol{l}}}, \quad (16)$$

where

$$p(\boldsymbol{l}) = \mathbb{P}\left\{ \sum_{i=1}^{I} \sum_{j=1}^{l_i} \zeta_i^{(j)} \geq x \;\middle|\; \zeta_i^{(j)} > 0 \text{ for all } i \text{ and } j = 1, \ldots, l_i \right\}$$

$$= \mathbb{P}\left\{ \sum_{i=1}^{I} \sum_{j=1}^{l_i} \xi_i^{(j)} \geq x \right\}$$

with the random variables $\xi_i^{(j)}$ defined in Theorem 1.

Now, on the event $\{\nu = N\}$,

$$\frac{\boldsymbol{Q}^{(N+1+n)}}{\rho^n} \to \zeta_N \boldsymbol{v} \quad \text{a.s. as } n \to \infty,$$

where

$$\mathbb{P}\{\zeta_N \in \cdot\} = \mathbb{P}\{\zeta_G \in \cdot \mid \boldsymbol{Z}^{(n)} \neq \boldsymbol{0} \text{ for all } n \in \mathbb{Z}_+\} = \mathbb{P}\{\zeta_G \in \cdot \mid \zeta_G > 0\}.$$

Then

$$\frac{\boldsymbol{Q}(t^{(n)})}{\rho^n} \to \underbrace{\left(\sum_{N\in\mathbb{Z}_+} 1(\nu = N)\frac{\zeta_N}{\rho^{N+1}}\right)}_{=:\,\zeta}\boldsymbol{v} \quad \text{a.s. as } n \to \infty,$$

and it is left to check that the distribution of ζ is given by (15).

For $x \in (0, \infty)$, we have

$$\mathbb{P}\{\zeta \geq x\} = \sum_{N\in\mathbb{Z}_+} \frac{\mathbb{P}\{\nu = N\}\mathbb{P}\{\zeta_G \geq \rho^{N+1}x\}}{\mathbb{P}\{\zeta_G > 0\}},$$

and then (15) follows by (16).

To deal with the other $t_i^{(n)}$, we combine the previous lemma with laws of large numbers.

Lemma 7. *For $i = 1, \ldots, I$, there exist constants $b_i \in (0, \infty)$ and $\boldsymbol{a}_i = (a_{i,1}, \ldots, a_{i,I}) \in [0, \infty)^I$ such that*

$$\frac{t_i^{(n)}}{\rho^n} \to b_i\zeta \quad \textit{and} \quad \boldsymbol{Q}\left(\frac{t_i^{(n)}}{\rho^n}\right) \to \zeta\boldsymbol{a}_i \quad \textit{a.s. as } n \to \infty.$$

The b_i and $\boldsymbol{a}_i$ are given by

$$b_1 = \frac{\sum_{i=1}^I v_i/\mu_i}{\sum_{i=1}^I \lambda_i/\mu_i - 1}, \qquad b_{i+1} = b_i + (v_i + \lambda_i(b_i - b_1))\gamma_i, \qquad i = 1, \ldots, I, \tag{17}$$

and

$$\boldsymbol{a}_1 = \boldsymbol{v}, \qquad \boldsymbol{a}_{i+1} = \boldsymbol{a}_i + (b_{i+1} - b_i)\boldsymbol{\lambda} - (b_{i+1} - b_i)\mu_i\boldsymbol{e}_i, \qquad i = 1, \ldots, I. \tag{18}$$

The $\boldsymbol{a}_i$ also satisfy

$$\boldsymbol{a}_1 = \boldsymbol{v}, \qquad \boldsymbol{a}_{i+1} = \boldsymbol{a}_i - a_{i,i}\boldsymbol{e}_i + a_{i,i}\check{\boldsymbol{m}}_i, \qquad i = 1, \ldots, I. \tag{19}$$

Remark 6. As we compare (17) and (18) with (3) and (4), it immediately follows that

$$b_i = \alpha\overline{b}_i, \qquad \boldsymbol{a}_i = \alpha\overline{\boldsymbol{a}}_i$$

Proof of Lemma 7. First we show that the sequences of $t_i^{(n)}/\rho^n$ and $\boldsymbol{Q}(t_i^{(n)})$ converge a.s., and that their limits satisfy the relations (18). Then we derive equations (17) and (19) relying on an LLN that, generally speaking, guarantees the b_i to be in-probability limits only.

6.2.1. *Asymptotics of* $t_1^{(n)}$. By the definition of ν, which is a.s. finite, we have, for $n > \nu$,

$$t_1^{(n)} = t^{(n)} = \underbrace{\sum_{l=0}^{\nu} (t_1^{(l)} - t^{(l)})}_{=:\,\Sigma} + \sum_{i=1}^{I} \sum_{k=1}^{D_i(t^{(n)})} B_i^{(k)}.$$

where Σ is a.s. finite.

The last equation, with $D_i(t^{(n)}) = E_i(t^{(n)}) - Q_i(t^{(n)})$ substituted, can be transformed into

$$t_1^{(n)} = t^{(n)} = \Sigma + t^{(n)} \underbrace{\sum_{i=1}^{I} \frac{\sum_{k=1}^{D_i(t^{(n)})} B_i^{(k)}}{D_i(t^{(n)})} \frac{E_i(t^{(n)})}{t^{(n)}}}_{=:\,\Sigma_1^{(n)}} - \rho^n \underbrace{\sum_{i=1}^{I} \frac{\sum_{k=1}^{D_i(t^{(n)})} B_i^{(k)}}{D_i(t^{(n)})} \frac{Q_i(t^{(n)})}{\rho^n}}_{=:\,\Sigma_2^{(n)}},$$

and then into

$$\frac{t_1^{(n)}}{\rho^n} = \frac{t^{(n)}}{\rho^n} = \frac{\Sigma_2^{(n)} - \Sigma/\rho^n}{\Sigma_1^{(n)} - 1}. \tag{20}$$

By the SLLN and Lemma 6,

$$\Sigma_1^{(n)} \to \sum_{i=1}^{I} \frac{\lambda_i}{\mu_i} \quad \text{and} \quad \Sigma_2^{(n)} \to \left(\sum_{i=1}^{I} \frac{v_i}{\mu_i} \right) \zeta \quad \text{a.s. as } n \to \infty,$$

which, together with (20), implies that

$$\frac{t^{(n)}}{\rho^n} \to b_1 \zeta \quad \text{and} \quad \frac{t_1^{(n)}}{\rho^n} \to b_1 \zeta \quad \text{a.s. as } n \to \infty, \tag{21}$$

where the value of b_1 is the one claimed in the lemma.

6.2.2. *Convergence of* $t_i^{(n)}/\rho^n$. Note that

$$t_{i+1}^{(n)} - t_i^{(n)} = I_i(t^{(n+1)}) - I_i(t^{(n)}),$$

and, hence,

$$\frac{t_{i+1}^{(n)}}{\rho^n} = \frac{t_i^{(n)}}{\rho^n} + \frac{I_i(t^{(n+1)})}{B_i(I_i(t^{(n+1)}))} \frac{D_i(t^{(n+1)})}{\rho^{n+1}} \rho - \frac{I_i(t^{(n)})}{B_i(I_i(t^{(n)}))} \frac{D_i(t^{(n)})}{\rho^n}. \tag{22}$$

By the SLLN,

$$\frac{B_i(I_i(t^{(n)}))}{I_i(t^{(n)})} \to \mu_i \quad \text{a.s. as } n \to \infty. \tag{23}$$

By the SLLN, (20) and Lemma 6,

$$\frac{D_i(t^{(n)})}{\rho^n} = \frac{E_i(t^{(n)})}{t^{(n)}} \frac{t^{(n)}}{\rho^n} - \frac{Q_i(t^{(n)})}{\rho^n} \to (\lambda_i b_1 - v_i)\zeta \quad \text{a.s. as } n \to \infty. \tag{24}$$

As we put (21)–(24) together, it follows that there exist positive numbers b_i such that

$$\frac{t_i^{(n)}}{\rho^n} \to b_i \zeta \quad \text{a.s. as } n \to \infty, \qquad i = 1, \ldots, I+1. \tag{25}$$

(The value of b_1 is the one claimed in the lemma, and the equations for the other b_i that follow from (21)–(24) are not given here since they will not be used anywhere in the proofs.)

6.2.3. *Convergence of $Q(t_i)/\rho^n$ and (18).* Since, during the time interval $[t_i^{(n)}, t_{i+1}^{(n)})$, there are no departures from queues other than i, we have

$$Q_j(t_{i+1}^{(n)}) - Q_j(t_i^{(n)}) = E_j(t_{i+1}^{(n)}) - E_j(t_i^{(n)}) - 1(j=i)(B_i(I_i(t^{(n+1)})) - B_i(I_i(t^{(n)}))). \quad (26)$$

By the SLLN and (25),

$$\frac{E_j(t_{i+1}^{(n)}) - E_j(t_i^{(n)})}{\rho^n} \to \lambda_j(b_{i+1} - b_i)\zeta \quad \text{a.s. as } n \to \infty. \quad (27)$$

By (25),

$$\frac{I_i(t^{(n)})}{\rho^n} = \frac{\sum_{k=1}^{n-1}(t_{i+1}^{(k)} - t_i^{(k)})}{\rho^n} \to \frac{b_{i+1} - b_i}{\rho - 1}\zeta \quad \text{a.s. as } n \to \infty, \quad (28)$$

which, together with the SLLN, implies that

$$\frac{B_i(I_i(t^{(n+1)})) - B_i(I_i(t^{(n)}))}{\rho^n} \to \mu_i(b_{i+1} - b_i)\zeta \quad \text{a.s. as } n \to \infty \quad (29)$$

As we put Lemma 6 and (26)–(29) together, it follows that

$$\boldsymbol{Q}(t_i^{(n)})/\rho^n \to \zeta \boldsymbol{a}_i \quad \text{a.s. as } n \to \infty, \qquad i = 1, \dots, I+1,$$

where the vectors $\boldsymbol{a}_i = (a_{i,1}, \dots, a_{i,I})$ are given by (18).

Proof of (17) and (19). We derive (17) from the equations

$$t_{i+1}^{(n)} = t_i^{(n)} + \sum_{k=1}^{Q_i(t_i^{(n)})} V_i^{(n,k)}, \quad (30)$$

$$Q_i(t_i^{(n)}) = Q_i(t_1^{(n)}) + E_i(t_i^{(n)}) - E_i(t_1^{(n)}). \quad (31)$$

To (30), we apply the following form of the LLN (the proof is postponed to Appendix A).

Proposition 2. *Let a random variable Y have a finite mean value and, for each $n \in \mathbb{N}$, let $Y_n^{(k)}$, $k \in \mathbb{N}$, be i.i.d. copies of Y. Let τ_n, $n \in \mathbb{N}$, be $\mathbb{Z}_+$-valued random variables such that τ_n is independent of the sequence $\{Y_n^{(k)}\}_{k\in\mathbb{N}}$ for each n and $\tau_n \to \infty$ in probability as $n \to \infty$. Finally, let a sequence $\{T_n\}_{n\in\mathbb{N}}$ of positive numbers increase to ∞. If there exists an a.s. finite random variable τ such that $\tau_n/T_n \to \tau$ in probability as $n \to \infty$, then*

$$\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{T_n} \to \tau \mathbb{E}Y \quad \textit{in probability as } n \to \infty.$$

By (30) and Proposition 2,

$$b_{i+1} - b_i = a_{i,i}\gamma_i. \quad (32)$$

By (31) and the SLLN,

$$a_{i,i} = v_i + \lambda_i(b_i - b_1). \quad (33)$$

Then (17) follows by substituting (33) into (32).

Finally, (19) follows as we apply Statement 2 to the equation

$$\boldsymbol{Q}(t_{i+1}^{(n)}) = \boldsymbol{Q}(t_i^{(n)}) - Q_i(t_i^{(n)})\boldsymbol{e}_i + \sum_{k=1}^{Q_i(t_i^{(n)})} \check{\boldsymbol{L}}_i^{(n,k)}.$$

6.3. Proof of Theorem 1

This proof converts the results of Lemma 7 using the following tool.

Lemma 8. *Suppose that random variables $Y^{(n)}$, $n \in \mathbb{Z}$, and Y are such that*

$$\frac{Y^{(n)}}{\rho^n} \to Y\zeta \quad \text{a.s. as } n \to \infty.$$

Then, for all $k \in \mathbb{Z}$,

$$\frac{Y^{(\eta_n+k)}}{\rho^n} \to Y\frac{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}\rho^k}{\alpha} \quad \text{a.s. as } n \to \infty.$$

Proof. First we show that, a.s.,

$$n - \eta_n = \lfloor \log_\rho(\alpha\zeta)\rfloor \quad \text{for all } n \text{ big enough.} \tag{34}$$

Indeed, we have $\log_\rho(t^{(n)}) - n = \log_\rho(\alpha\zeta) + \delta^{(n)}$, where $\delta^{(n)} \to 0$ a.s. as $n \to \infty$. Then $\eta_n = \min\{k\colon \log_\rho(t^{(k)}) \geq n\} = \min\{k\colon k + \delta^{(k)} \geq n - \log_\rho(\alpha\zeta)\}$. Introduce the event $\Omega' := \{\delta^{(n)} \to 0, \log_\rho(\alpha\zeta) \notin \mathbb{Z}\}$. When estimated at any $\omega \in \Omega'$, $\eta_n = \lceil n - \log_\rho(\alpha\zeta)\rceil = n - \lfloor \log_\rho(\alpha\zeta)\rfloor$ for all n big enough; and $\mathbb{P}\{\Omega'\} = 1$ since the distribution function of random variable ζ is continuous in $(0, \infty)$ (see (15), where the random variables $\xi_i^{(j)}$ have continuous densities on $(0, \infty)$ by Proposition 1).

Now fix a $k \in \mathbb{Z}$. By (34),

$$\frac{Y^{(\eta_n+k)}}{\rho^n} = \frac{Y^{(\eta_n+k)}}{\rho^{\eta_n+k}}\frac{\rho^k}{\rho^{n-\eta_n}} \to Y\zeta\frac{\rho^k}{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}} \quad \text{a.s. as } n \to \infty,$$

where

$$\frac{\zeta}{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}} = \frac{\rho^{\log_\rho(\alpha\zeta)}}{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}}\frac{1}{\alpha} = \frac{\rho^{\{\log_\rho(\alpha\zeta)\}}}{\alpha},$$

and, hence, Lemma 8 is proven.

Now we proceed with the proof of Theorem 1.

Lemmas 7 and 8 imply that the convergence (2) holds with

$$\xi := \rho^{\{\log_\rho(\alpha\zeta)\}}.$$

By definition, ξ takes values in $[1, \rho)$, and it is left to calculate its distribution.

Fix an $x \in [1, \rho)$. Since

$$\begin{aligned}\mathbb{P}\{\xi \geq x\} &= \mathbb{P}\{\{\log_\rho(\alpha\zeta)\} \geq \log_\rho x\}\\ &= \sum_{m\in\mathbb{Z}} \mathbb{P}\{m + \log_\rho x \leq \log_\rho(\alpha\zeta) < m + 1\}\\ &= \sum_{m\in\mathbb{Z}} \mathbb{P}\left\{\frac{\rho^m x}{\alpha} \leq \zeta < \frac{\rho^{m+1}}{\alpha}\right\},\end{aligned}$$

we have, by Lemma 6,

$$\mathbb{P}\{\xi \geq x\} = \frac{1}{1-q_G} \sum_{n\in\mathbb{Z}_+} \mathbb{P}\{v=n\} \sum_{\substack{\boldsymbol{k}\in\mathbb{Z}_+^I,\\ |\boldsymbol{k}|\geq 1}} G(\boldsymbol{k}) \sum_{\substack{\boldsymbol{l}\leq\boldsymbol{k},\\ |\boldsymbol{l}|\geq 1}} \binom{\boldsymbol{k}}{\boldsymbol{l}} (\mathbf{1}-\boldsymbol{q})^{\boldsymbol{l}} \boldsymbol{q}^{\boldsymbol{k}-\boldsymbol{l}}$$
$$\times \underbrace{\sum_{m\in\mathbb{Z}} \mathbb{P}\left\{ \frac{\rho^{m+n+1}x}{\alpha} \leq \sum_{i=1}^{I}\sum_{j=1}^{l_i} \xi_i^{(j)} < \frac{\rho^{m+n+2}}{\alpha} \right\}}_{=:\,\Sigma_{n,l}}.$$

Note that $\Sigma_{n,l}$ does not depend on n,

$$\begin{aligned}\Sigma_{n,l} &= \sum_{m\in\mathbb{Z}} \mathbb{P}\left\{ \frac{\rho^m x}{\alpha} \leq \sum_{i=1}^{I}\sum_{j=1}^{l_i} \xi_i^{(j)} < \frac{\rho^{m+1}}{\alpha} \right\} \\ &= \sum_{m\in\mathbb{Z}} \mathbb{P}\left\{ m + \log_\rho x \leq \log_\rho\left(\alpha \sum_{i=1}^{I}\sum_{j=1}^{l_i} \xi_i^{(j)} \right) < m+1 \right\} \\ &= \mathbb{P}\left\{ \left\{ \log_\rho\left(\alpha \sum_{i=1}^{I}\sum_{j=1}^{l_i} w_i^{(j)} \right) \right\} \geq \log_\rho x \right\},\end{aligned}$$

and this finishes the proof of Theorem 1.

6.4. Proof of Theorem 2

The proof consists of several steps. Throughout the proof, we assume that the function $\overline{\boldsymbol{Q}}(\cdot)$ is defined by (6). First we show that the process $\xi\,\overline{\boldsymbol{Q}}(\cdot/\xi)$ coincides a.s. with the pointwise limit of the scaled processes $\overline{\boldsymbol{Q}}^{(n)}(\cdot)$. Then we check that $\overline{\boldsymbol{Q}}(\cdot)$ satisfies (5) and is continuous. Finally, we prove that the pointwise convergence of the processes $\overline{\boldsymbol{Q}}^{(n)}(\cdot)$ implies their uniform convergence on compact sets.

6.4.1. *Pointwise convergence.* To start with, we define the auxiliary event Ω' on which, as $n\to\infty$,

$$\frac{t_i^{(\eta_n+k)}}{\rho^n} \to \rho^k \overline{b}_i \xi \quad \text{and} \quad \frac{\boldsymbol{Q}(t_i^{(\eta_n+k)})}{\rho^n} \to \xi\rho^k \overline{a}_i, \qquad i=1,\dots,I+1, \quad k\in\mathbb{Z},$$
$$\frac{I(t_i^{(\eta_n+k)})}{\rho^n} \to \rho^k \frac{\overline{b}_{i+1}-\overline{b}_i}{\rho(\rho-1)}\xi, \qquad i=1,\dots,I, \quad k\in\mathbb{Z},$$
$$\frac{E_i(t)}{t} \to \lambda_i \quad \text{and} \quad \frac{B_i(t)}{t} \to \mu_i, \qquad i=1,\dots,I.$$

By Theorem 1, (28) and the SLLN, $\mathbb{P}\{\Omega'\}=1$.

We will now show that on Ω', as $n\to\infty$,

$$\overline{\boldsymbol{Q}}^{(n)}(t) \to \xi\,\overline{\boldsymbol{Q}}\left(\frac{t}{\xi}\right) \quad \text{for all } t\in[0,\infty), \tag{35}$$

where $\overline{\boldsymbol{Q}}(\cdot)$ is given by (6).

Fix a queue number i and an outcome $\omega \in \Omega'$. All random objects in the rest of this part of the proof will be evaluated at this ω.

For $t = 0$, the convergence (35) holds since the system starts empty. For $t > 0$, we consider the three possible cases.

Case 1: $t \in [\rho^k \overline{b}_i \xi, \rho^k \overline{b}_{i+1} \xi)$ *for a* $k \in \mathbb{Z}$. By the definition of Ω', for all n big enough,

$$\frac{t_i^{(\eta_n+k)}}{\rho^n} < t < \frac{t_{i+1}^{(\eta_n+k)}}{\rho^n},$$

implying that queue i is in service during $[t_i^{(\eta_n+k)}, \rho^n t)$, and hence

$$Q_i(\rho^n t) = Q_i(t_i^{(\eta_n+k)}) + (E_i(\rho^n t) - E_i(t_i^{(\eta_n+k)}) - (D_i(\rho^n t) - D_i(t_i^{(\eta_n+k)}),$$

where

$$D_i(\rho^n t) - D_i(t_i^{(\eta_n+k)}) = B_i(I_i(t_i^{(\eta_n+k)}) + (\rho^n t - t_i^{(\eta_n+k)})) - B_i(I_i(t_i^{(\eta_n+k)})).$$

Again by the definition of Ω', the last two equations imply that, as $n \to \infty$,

$$\overline{Q}_i^{(n)}(t) \to \rho^k \overline{a}_{i,i} \xi + \lambda_i(t - \rho^k \overline{b}_i \xi) - \mu_i(t - \rho^k \overline{b}_i \xi) = \xi \overline{Q}_i(t/\xi).$$

Case 2: $t \in [\rho^k \overline{b}_{i+1} \xi, \rho^{k+1} \overline{b}_i \xi)$ *for a* $k \in \mathbb{Z}$. In this case, for all n big enough,

$$\frac{t_{i+1}^{(\eta_n+k)}}{\rho^n} < t < \frac{t_i^{(\eta_n+k+1)}}{\rho^n},$$

and, hence, queue i is not in service during $[\rho^n t, t_i^{(\eta_n+k+1)})$, i.e.

$$Q_i(t_i^{(\eta_n+k+1)}) = Q_i(\rho^n t) + E_i(t_i^{(\eta_n+k+1)}) - E_i(\rho^n),$$

implying that

$$\overline{Q}_i^{(n)}(t) \to \rho^{k+1} a_{i,i} \xi - \lambda_i(\rho^{k+1} \overline{b}_i \xi - t) = \xi \overline{Q}_i\left(\frac{t}{\xi}\right).$$

Case 3: $t = \rho^k \overline{b}_i \xi$ *for a* $k \in \mathbb{Z}$. Since, as $n \to \infty$,

$$\frac{t_{i+1}^{(\eta_n+k-1)}}{\rho^n} \to \rho^{k-1} \overline{b}_{i+1} \xi, \qquad \frac{t_i^{(\eta_n+k)}}{\rho^n} \to t \quad \text{and} \quad \frac{t_{i+1}^{(\eta_n+k)}}{\rho^n} \to \rho^k \overline{b}_{i+1} \xi$$

and the limits satisfy the inequality

$$\rho^{k-1} \overline{b}_{i+1} \xi < t < \rho^k \overline{b}_{i+1} \xi,$$

all n big enough fall into the two sets

$$\mathcal{N}_1 := \{n : t_i^{(\eta_n+k)} \le \rho^n t < t_{i+1}^{(\eta_n+k)}\} \quad \text{and} \quad \mathcal{N}_2 := \{n : t_{i+1}^{(\eta_n+k-1)} < \rho^n t < t_i^{(\eta_n+k)}\}.$$

For $l = 1, 2$, we have to check that, if the set $\mathcal{N}_l$ is infinite, then

$$\overline{Q}_i^{(n)}(t) \to \rho^k \overline{a}_{i,i} \xi \quad \text{as } n \to \infty, n \in \mathcal{N}_l. \tag{36}$$

For $l = 1$, (36) follows along the lines of Case 1. For $l = 2$, we can prove (36) following the lines of Case 2 and replacing $k + 1$ with k.

6.4.2. *Equivalence of (5) and (6).* Let $\widetilde{\boldsymbol{Q}}(\cdot) = (\widetilde{Q}_1, \dots, \widetilde{Q}_I)(\cdot)$ be the unique solution to (5), whereas $\overline{\boldsymbol{Q}}(\cdot)$, as before, is given by (5). Fix a queue number i. The slopes of $\overline{Q}_i(\cdot)$ and $\widetilde{Q}_i(\cdot)$ coincide everywhere. Also $\overline{Q}_i(0) = 0 = \widetilde{Q}_i(0)$, and $\overline{Q}_i(\rho^k \overline{b}_i) = \rho^k \overline{a}_{i,i} = \widetilde{Q}_i(\rho^k \overline{b}_i)$, $k \in \mathbb{Z}$. Then it is left to check that

$$\overline{Q}_i(\rho^k \overline{b}_j) = \rho^k \overline{a}_{j,i} = \widetilde{Q}_i(\rho^k \overline{b}_j), \qquad j \neq i, \quad k \in \mathbb{Z}. \tag{37}$$

We have,

$$\begin{aligned} \rho^k \overline{b}_j \in [\rho^{k-1} \overline{b}_{i+1}, \rho^k \overline{b}_i) \quad \text{and} \quad \overline{Q}_i(\rho^k \overline{b}_j) = \rho^k (\overline{b}_i - \lambda_i (\overline{b}_i - \overline{b}_j)), \qquad j < i, \\ \rho^k \overline{b}_j \in [\rho^k \overline{b}_{i+1}, \rho^{k+1} \overline{b}_i) \quad \text{and} \quad \overline{Q}_i(\rho^k \overline{b}_j) = \rho^k (\rho \overline{b}_i - \lambda_i (\rho \overline{b}_i - \overline{b}_j)), \qquad j > i. \end{aligned}$$

Then (37) follows from the equations

$$\begin{aligned} Q_i(t_i^{(n)}) = Q_i(t_j^{(n)}) + E_i(t_i^{(n)}) - E_i(t_j^{(n)}), \qquad j < i, \\ Q_i(t_i^{(n+1)}) = Q_i(t_j^{(n)}) + E_i(t_i^{(n+1)}) - E_i(t_j^{(n)}), \qquad j > i, \end{aligned}$$

by Lemma 7 and Remark 6.

6.4.3. *Continuity of* $\overline{Q}(\cdot)$. Fix a queue number i. As defined by (6), the function $\overline{Q}_i(\cdot)$ might have discontinuities only at $t = 0$ and $t = \rho^k \overline{b}_{i+1}$, $k \in \mathbb{Z}$.

Note that $\sup_{t \in [\rho^{k-1} \overline{b}_i, \rho^k \overline{b}_i)} \overline{Q}_i(t) = \rho^k \overline{a}_{i,i}$, $k \in \mathbb{Z}$. Then

$$\sup_{t \in (0, \rho^k \overline{b}_i)} \overline{Q}_i(t) = \sup_{l \in \mathbb{Z}, l \leq k} \rho^k \overline{a}_{i,i} \to 0 \quad \text{as } k \to -\infty,$$

and, hence, $\overline{Q}_i(t) \to 0 = \overline{Q}_i(0)$ as $t \to 0$.

At $t = \rho^k \overline{b}_{i+1}$, $k \in \mathbb{Z}$, the function $\overline{Q}_i(\cdot)$ is right-continuous with the left limit given by $\lim_{t \uparrow \rho^k \overline{b}_{i+1}} \overline{Q}_i(t) = \rho^k (\overline{a}_{i,i} + (\lambda_i - \mu_i)(\overline{b}_{i+1} - \overline{b}_i))$. By (4) and (37), we have

$$\lim_{t \uparrow \rho^k \overline{b}_{i+1}} \overline{Q}_i(t) = Q_i(\rho^k \overline{b}_{i+1}) = \rho^k a_{i+1,i}.$$

6.4.4. *Uniform convergence on compact sets.* Define the auxiliary event Ω'' on which, as $n \to \infty$, $\overline{\boldsymbol{Q}}^{(n)}(\cdot) \to \xi \overline{\boldsymbol{Q}}(\cdot/\xi)$ pointwise, and $E_i(\rho^n \cdot)/\rho^n \to \lambda_i \cdot$ uniformly on compact sets, $i = 1, \dots, I$. As follows from the first part of the proof and the functional SLLN, $\mathbb{P}\{\Omega''\} = 1$. For the rest of the proof, we estimate random objects at an outcome $\omega \in \Omega''$. Consider the scaled departure processes $D_i(\rho^n \cdot)/\rho^n = E_i(\rho^n \cdot)/\rho^n - \overline{Q}_i^{(n)}(\cdot)$. These processes are monotone and, by the definition of Ω'', converge pointwise to the continuous functions $\lambda_i \cdot - \xi \overline{Q}_i(\cdot/\xi)$. Then they converge uniformly on compact sets, and the same is true for the processes $\overline{Q}_i^{(n)}(\cdot)$.

Appendix A

Proof of Lemma 2. Suppose that $\rho \leq 1$. Then, by [8, Theorem 7.1], we have $q_i = 1$ for all i and $q_G = 1$. The latter implies that the queue length process $\boldsymbol{Q}(\cdot)$ hits $\mathbf{0}$ infinitely many times, and the same holds for the workload process. Let $\{t^{(n_k)}\}_{k \in \mathbb{Z}_+}$ be the sequence of consecutive time instants such that $\boldsymbol{Q}(t^{(n_k)}) = \mathbf{0}$. For different k, the differences $t^{(n_{k+1})} - t^{(n_k)}$ are bounded from below by the waiting times until the first arrival into the empty system, which are i.i.d. random variables distributed exponentially with parameter $\sum_{i=1}^I \lambda_i$. Hence, we

have $t^{(n_k)} \to \infty$ a.s. as $k \to \infty$. This leads to a contradiction with the fact that the system is overloaded and its total workload grows infinitely large with time (by the SLLN, $(\sum_{i=1}^{I} \sum_{k=1}^{E_i(t)} B_i^{(k)} - t)/t \to \sum_{i=1}^{I} \lambda_i/\mu_i - 1 > 0$ a.s. as $t \to \infty$). Hence, $\rho > 1$ and, consequently, [8, Theorem 7.1] implies that $q_i < 1$ for all i and $q_G < 1$.

Proof of Proposition 2. First we show that

$$\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n} \to \mathbb{E}Y \quad \text{in probability as } n \to \infty. \tag{38}$$

By the independence between τ_n and $\{Y_n^{(k)}\}_{k\in\mathbb{N}}$, for all $N \in \mathbb{Z}_+$,

$$\mathbb{P}\left\{\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n} - \mathbb{E}Y\right| \geq \varepsilon, \tau_n = N\right\} = \mathbb{P}\left\{\left|\sum_{k=1}^{N} \frac{Y_1^{(k)}}{N} - \mathbb{E}Y\right| \geq \varepsilon\right\}\mathbb{P}\{\tau_n = N\}.$$

Then, for any $M \in \mathbb{Z}_+$,

$$\mathbb{P}\left\{\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n} - \mathbb{E}Y\right| \geq \varepsilon\right\} \leq \mathbb{P}\{\tau_n \leq M\} + \sup_{N>M} \mathbb{P}\left\{\left|\sum_{k=1}^{N} \frac{Y_1^{(k)}}{N} - \mathbb{E}Y\right| \geq \varepsilon\right\},$$

and (38) follows as we first let $n \to \infty$, and then $M \to \infty$.

Now that we have shown (38), the statement follows by

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{T_n} - \tau\mathbb{E}Y\right| \geq \varepsilon\right\} \\
&\quad\leq \mathbb{P}\left\{\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n}\right|\left|\frac{\tau_n}{T_n} - \tau\right| \geq \frac{\varepsilon}{2}\right\} + \mathbb{P}\left\{\tau\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n} - \mathbb{E}Y\right| \geq \frac{\varepsilon}{2}\right\} \\
&\quad\leq \mathbb{P}\left\{C_1\left|\frac{\tau_n}{T_n} - \tau\right| \geq \frac{\varepsilon}{2}\right\} + \mathbb{P}\left\{\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n}\right| \geq C_1\right\} \\
&\qquad + \mathbb{P}\left\{C_2\left|\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{\tau_n} - \mathbb{E}Y\right| \geq \frac{\varepsilon}{2}\right\} + \mathbb{P}\{\tau \geq C_2\}
\end{aligned}$$

as we first let $n \to \infty$, and then $C_1 \to \infty$, $C_2 \to \infty$.

Acknowledgement

Sergey Foss thanks the Institute for Mathematical Research (FIM) at ETH Zurich for their partial support.

References

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

[2] Borst, S. C. (1996). *Polling Systems.* CWI, Amsterdam.

[3] Boxma, O. J. (1991). Analysis and optimization of polling systems. In *Queueing, Performance and Control in ATM*, eds J. W. Cohen and C. D. Pack. North-Holland, Amsterdam, pp. 173–183.

[4] Coffman, E. G., Jr., Puhalskii, A. A. and Reiman, M. I. (1995). Polling systems with zero switchover times: a heavy-traffic averaging principle. *Ann. Appl. Prob.* **5,** 681–719.
[5] Coffman, E. G., Jr., Puhalskii, A. A. and Reiman, M. I. (1998). Polling systems in heavy traffic: a Bessel process limit. *Math. Operat. Res.* **23,** 257–304.
[6] Foss, S. (1984). Queues with customers of several types. In *Advances in Probability Theory: Limit Theorems and Related Problems*, ed. A. A. Borovkov. Springer, New York, pp. 348–377.
[7] Foss, S. and Kovalevskii, A. (1999). A stability criterion via fluid limits and its application to a polling model. *Queueing Systems Theory Appl.* **32,** 131–168.
[8] Harris, T. E. (1963). *The Theory of Branching Processes.* Springer, Berlin.
[9] Kesten, H. and Stigum, B. P. (1966). A limit theorem for multidimensional Galton–Watson processes. *Ann. Math. Statist.* **37,** 1211–1223.
[10] Kovalevskiĭ, A. P., Topchii, V. A. and Foss, S. G. (2005). On the stability of a queueing system with continually branching fluid limits. *Problems Inf. Trans.* **41,** 254–279.
[11] Kroese, D. P. (1997). Heavy traffic analysis for continuous polling models. *J. Appl. Prob.* **34,** 720–732.
[12] Kurtz, T., Lyons, R., Pemantle, R. and Peres, Y. (1997). A conceptual proof of the Kesten–Stigum theorem for multi-type branching processes. In *Classical and Modern Branching Processes* (IMA Vol. Math. Appl. **84**), Springer, New York, pp. 181–185.
[13] Mack, C., Murphy, T. and Webb, N. L. (1957). The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants. *J. R. Statist. Soc. B* **19,** 166–172.
[14] MacPhee, I., Menshikov, M., Petritis, D. and Popov, S. (2007). A Markov chain model of a polling system with parameter regeneration. *Ann. Appl. Prob.* **17,** 1447–1473.
[15] MacPhee, I., Menshikov, M., Petritis, D. and Popov, S. (2008). Polling systems with parameter regeneration, the general case. *Ann. Appl. Prob.* **18,** 2131–2155.
[16] Resing, J. A. C. (1993). Polling systems and multitype branching processes. *Queueing Systems Theory Appl.* **13,** 409–426.
[17] Takagi, H. (1986). *Analysis of Polling Systems.* MIT Press, Cambridge, MA.
[18] Takagi, H. (1990). Queueing analysis of polling systems: an update. In *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi, North-Holland, Amsterdam, pp. 267–318.
[19] Takagi, H. (1997). Queueing analysis of polling models: progress in 1990–1994. In *Frontiers in Queueing*, CRC, Boca Raton, FL, pp. 119–146.
[20] Vatutin, V. A. and Dyakonova, E. E. (2002). Multitype branching processes and some queueing systems. *J. Math. Sci. (New York)* **111,** 3901–3911.
[21] Van der Mei, R. D. (2007). Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Systems* **57,** 29–46.
[22] Van der Mei, R. D. and Resing, J. A. C. (2007). Polling models with two-stage gated service: fairness versus efficiency. In *Managing Traffic Performance in Converged Networks* (Lecture Notes Comput. Sci. **4516**), Springer, Berlin, pp. 544–555.
[23] Van der Mei, R. D. and Roubos, A. (2012). Polling models with multi-phase gated service. *Ann. Operat. Res.* **198,** 25–56.
[24] Van Wijk, A. C. C., Adan, I. J. B. F., Boxma, O. J. and Wierman, A. (2012). Fairness and efficiency for polling models with the κ-gated service discipline. *Performance Evaluation* **69,** 274–288.
[25] Yechiali, U. (1993). Analysis and control of polling systems. In *Performance Evaluation of Computer and Communication Systems* (Lecture Notes Comput. Sci. **729**), Springer, Berlin, pp. 630–650.