EURANDOM PREPRINT SERIES

2017-001

January 20, 2017

Two queues with random time-limited polling

M. Saxena, O. Boxma, S. Kapodistria, R. Núñez Queija

ISSN 1389-2355

Two queues with random time-limited polling

Mayank Saxena*[†] Onno Boxma*[‡] Stella Kapodistria*[§] Rudesindo Núñez Queija [¶][‖]

January 20, 2017

This paper is dedicated to Tomasz Rolski, in friendship, respect and admiration. His love of applied probability and never-ending curiosity are a blessing for our field.

Abstract

In this paper, we analyse a single server polling model with two queues. Customers arrive at the two queues according to two independent Poisson processes. There is a single server that serves both queues with generally distributed service times. The server spends an exponentially distributed amount of time in each queue. After the completion of this residing time, the server instantaneously switches to the other queue, i.e., there is no switch-over time. For this polling model we derive the steady-state marginal workload distribution, as well as heavy traffic and heavy tail asymptotic results. Furthermore, we also calculate the joint queue length distribution for the special case of exponentially distributed service times using singular perturbation analysis.

Keywords: polling model; workload decomposition; heavy traffic; heavy tail asymptotics; singular perturbation analysis; time-scale separation; geometric ergodicity

1 Introduction

In this paper, we are interested in the performance analysis of a single server polling model with a special service discipline (i.e., the criterion which determines how many customers are served during a visit of the server to a queue). A typical polling model consists of multiple queues, attended by a single server who visits the queues in some order to render service to the customers waiting at the queues. Moving from one queue to another, the server incurs a (possibly zero) switch-over time. Once the server is at one of the queues, the server serves the customers of that queue based on a service discipline and according to some service time distribution.

Polling models were initially introduced in the 1950's but mostly gained their popularity during the 1990's. This popularity rise was due to the wide range of applicability of polling models, especially for the modelling of computer-communication systems and protocols, traffic signal management, and manufacturing, see, e.g., [34, 35, 38] for a series of comprehensive surveys and [5, 26, 33] for extensive overviews of the applicability of polling systems.

The performance analysis of polling models has received considerable attention, see, e.g., [32]. In particular, in the polling literature much attention has been given to determining the probability generating function (PGF)

*Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands

[†] Email: m.mayank@tue.nl

[‡] Email: o.j.boxma@tue.nl

[§] Email: s.kapodistria@tue.nl

[¶] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands, Email: nunezqueija@uva.nl

[‖] CWI, Amsterdam, The Netherlands

of the joint queue length distribution under stationarity and at various epochs. A wide range of service disciplines has been considered, including *exhaustive service* (per visit to a queue, the server continues to serve all customers until it empties) and *gated service* (per visit to a queue, the server serves only those customers which are already present at the start of the visit). In [29], Resing shows that the joint queue length PGF of polling models in which the service discipline satisfies the so-called branching property equals the (known) PGF of a multi-type branching process with immigration. Service disciplines which satisfy the branching property include the exhaustive and gated disciplines. Polling systems with disciplines which do not satisfy the branching property usually defy an exact analysis. In our paper, we assume that the server spends an exponentially distributed amount of time at each queue. Upon the completion of this residing time at each queue, the server instantaneously switches to another queue according to a cyclic order. Such a service protocol does not exhibit the branching property, which complicates the analysis significantly. We concentrate on the two-queue model and, whenever possible, suggest extensions to the multi-queue model. A similar service discipline has been considered in [16], [40], [14], and the references therein.

In [16], the authors consider a multi-queue polling system under the Randomly Timed Gated (RTG) service discipline. The RTG discipline operates as follows: whenever the server enters a station, a timer is activated. If the server empties the queue before the timer's expiration, the server moves on to the next queue. Otherwise (i.e., if there is still work in the station when the timer expires), the server obeys one of the following rules, each leading to a different model: (1) The server completes all the work accumulated up to the timer's expiration and then moves on to the next node. (2) The server completes only the service of the job currently being served, and moves on. (3) The server stops working immediately and moves on. The model suggested in this manuscript bears resemblance to rule (3), however, in our case if a queue becomes empty, the server does not switch, and only does so when the timer expires.

In [40], the authors consider a single server multi-queue system, in which the server visits the individual queues for a fixed amount of time in a deterministic, cyclic order. Xie et al. [40] refer to the fixed residing time as the *orientation* time. They argue that such a service discipline comes with two operational advantages: it enables to i) keep the frequency of switching at a predetermined level (thus controlling the total cost, if there is a switching cost), ii) balance the time that the server spends in each queue (since, contrary to exhaustive or gated service disciplines, this discipline does not depend on the number of customers present in the various queues).

In [14], the authors assume a random visit (residing) time for each queue, which is independent of the number of customers present at each queue, and a *preemptive-repeat* with resampling service strategy. This autonomous service discipline is motivated from application in wireless ad hoc networks with movable communication hops. Another application is in single upstream tree-based ethernet passive optical networks, in which the central optical line terminal dedicates the channel to a specific user (e.g., the user with the highest priority) for a random amount of time, see, [24] and the references therein. For more applications on this type of autonomous service disciplines, the interested reader is referred to [1]. For all aforementioned applications, we consider it natural to assume that the service strategy is *preemptive-resume* and that the switch-over time is negligible in comparison to the service time and the residing time.

In this paper, we also devote attention to the individual queues. When focusing on a single queue, the model can be interpreted as a service system with *vacations*: we interpret the time that the server visits the other queue as a vacation period. Vacation queues - and *priority* queues for which the mathematical analysis is similar - are well studied in the queueing literature starting with the work of White and Christie [39] (exponentially distributed service times and vacations), Gaver[22], Thiruvengadam [37] and Avi-Itzhak and Naor [3] (the latter three assuming generally distributed service times and vacations). All these works assume that the service periods have an exponential distribution, but vary for example in the assumptions regarding whether interrupted services are resumed or repeated and in the metrics of interest. Takagi [34] provides an excellent overview of vacation and priority systems. The interested reader is also referred to Federgruen and Green [18] for phase-type distributed service periods, to Takine and Sengupta [36] for Markovian arrival processes and to Fiems et al. [19] for a more recent publication with various sorts of service disruptions. For a more extensive overview of the literature, we refer to the recent survey of Krishnamoorthy et al. [25].

A particular feature of a large class of vacation queues is that the stationary workload and queue length distributions obey a stochastic decomposition property, as first observed by Gaver [22] and Miller [27]. Fuhrmann and Cooper [20] give conditions for such a queue length decomposition to hold. Our model does not satisfy these conditions, but we show that it does allow a stochastic decomposition of the stationary workload.

The paper is organized as follows: In Section 2, we describe the two-dimensional polling model under consideration. In Section 3, we present the Laplace-Stieltjes transforms (LSTs) of the model's *marginal* workload distributions in steady-state at an arbitrary epoch. In Section 4, we show that a single queue's marginal workload satisfies a decomposition property. In Section 5, we derive the heavy traffic and heavy tail asymptotics of the marginal workload distributions in steady-state. We then discuss, in Section 6, open problems arising in the calculation of the *joint* workload distribution. Assuming exponentially distributed service times, we calculate in Section 7 the joint queue length distribution in steady-state at an arbitrary epoch using singular perturbation analysis. Several possible future research directions are discussed in Section 8.

2 Model description and notation

In this paper, we consider a two-queue polling model. Customers arrive to queue i according to a Poisson process at rate λ_i, $i = 1,2$. There is a single server, that serves both queues according to the first come first serve (FCFS) discipline. The service times of customers in queue i are independent and identically generally distributed positive random variables, say B_i, $i = 1,2$. We denote the LST of the service time B_i by $\tilde{b}_i(s) = \mathbb{E}(e^{-sB_i})$, with Re $s \geq 0$, $i = 1,2$.

A special feature of the polling model under consideration is that the server spends an exponentially distributed amount of time at queue i with rate c_i, $i = 1,2$. Upon completion of the residing time at queue i, the server instantaneously switches to the other queue, i.e., there is no switch-over time. Furthermore, if upon completion of the residing time, the server is providing service to a customer, this service is interrupted and resumed at the next visit of the server to the queue. More explicitly, we assume that if a server resumes the service after being interrupted, the server continues from where the service stopped instead of starting from the beginning, i.e., the service is *preemptive–resume*. Let T_i denote the residing time of the server in queue i, with T_i exponentially distributed with rate c_i and probability density function $f_{T_i}(t) = c_i e^{-c_i t}$, $t \geq 0$, $i = 1,2$. We denote the LST of the residing time T_i by $\tilde{f}_{T_i}(s) = \mathbb{E}(e^{-sT_i})$, with Re $s \geq 0$, $i = 1,2$.

Stability condition. For the two-queue polling model under consideration the stability condition is

$$\rho_1 < \frac{c_2}{c_1 + c_2} \text{ and } \rho_2 < \frac{c_1}{c_1 + c_2}, \tag{2.1}$$

with $\rho_i = \lambda_i \mathbb{E}(B_i)$, $i = 1,2$. The stability condition for the first queue can be easily proven by showing that the long-run proportion of time the server spends in the first queue is equal to $c_2/(c_1 + c_2)$, thus the long-run rate of service in the first queue is $c_2/(c_1 + c_2)\mathbb{E}(B_1)$. Hence, for the first queue to be stable it is needed that the arrival rate is strictly smaller than the long-run rate of service, which proves the left hand side of (2.1). The stability condition for the second queue can be proven in an analogous manner.

3 Marginal workload analysis

In this section, we derive the distribution of the marginal workload in steady-state at an arbitrary epoch. As discussed in the introductory section, the individual queues behave as vacation systems: from the perspective of one queue, the server is on vacation when it resides at the other queue. In this section, we give a direct derivation of the stationary marginal workload distributions.

We let $V_i(t)$ denote the workload at time t, $t \geq 0$, of queue i, $i = 1,2$, and V_i denote the steady-state workload of queue i at an arbitrary epoch, $i = 1,2$.

Theorem 3.1. *The LST of the workload of the first queue in steady-state under the stability condition* (2.1) *is given by*

$$\mathbb{E}(e^{-sV_1}) = \frac{s\left[\lambda_1\mathbb{E}(B_1)(c_1+c_2)-c_2\right]\left[c_1+c_2+\lambda_1\left(1-\tilde{b}_1(s)\right)\right]}{\left[\left(c_2+\lambda_1\left(1-\tilde{b}_1(s)\right)\right)\left(c_1+\lambda_1\left(1-\tilde{b}_1(s)\right)-s\right)-c_1c_2\right](c_1+c_2)}. \tag{3.1}$$

A symmetric formula holds for the LST of V_2 under the stability condition (2.1).

Proof. The derivation of the LST of the steady-state workload for the first queue is performed by considering the renewal process at the instances the server arrives at the first queue, i.e., the inter-renewal times are identical in distribution to T_1+T_2, with $T_i \sim \text{Exp}(c_i)$, $i=1,2$.

To structure the exposition, the proof of the theorem is split into five steps. A key point of the proof is the derivation of $\mathbb{E}(e^{-sV_1(T_1+T_2)})$; this is achieved in Step 4, after we derive $\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(T_1)=y)$ in Step 1, $\mathbb{E}(e^{-sV_1(T_1)}|V_1(0)=v)$ in Step 2, and subsequently $\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(0)=v)$ in Step 3. Finally, in Step 5, we calculate $\mathbb{E}(e^{-sV_1})$ using the PASTA property and the result of Step 4.

Step 1: Calculation of $\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(T_1)=y)$.

During $(T_1, T_1+T_2]$ the server serves only customers in the second queue, so the workload in the first queue only increases by the sum of the service times of all the customers that arrived within this interval. The increments occur according to a compound Poisson process. So,

$$\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(T_1)=y) = e^{-sy}\tilde{f}_{T_2}(\lambda_1(1-\tilde{b}_1(s))) = e^{-sy}\frac{c_2}{c_2+\lambda_1(1-\tilde{b}_1(s))}. \tag{3.2}$$

Step 2: Calculation of $\mathbb{E}(e^{-sV_1(T_1)}|V_1(0)=v)$.

Note that

$$\mathbb{E}(\mathrm{e}^{-sV_1(T_1)}|V_1(0)=v) = \int_{t=0}^{\infty} c_1\mathrm{e}^{-c_1t}\int_{\sigma=0}^{\infty} e^{-s\sigma}\mathrm{d}\mathbb{P}(V_1(t)<\sigma|V_1(0)=v)\mathrm{d}t. \tag{3.3}$$

In order to calculate the right hand side of (3.3), we use [11, p. 262, Equation (4.99)]

$$\begin{aligned}\int_{\sigma=0}^{\infty} e^{-s\sigma}\mathrm{d}\mathbb{P}(V_1(t)<\sigma|V_1(0)=v) &= e^{s(t-v)-t\lambda_1(1-\tilde{b}_1(s))}\\ &\quad - sU_1(t-v)\int_{u=0}^{t-v} e^{(s-\lambda_1(1-\tilde{b}_1(s)))(t-u-v)}\mathbb{P}(V_1(u+v)=0|V_1(0)=v)\mathrm{d}u,\end{aligned}$$

with Re $s\geq 0$, $t\geq 0$, and $U_1(x)=0$, if $x<0$, and $U_1(x)=1$, otherwise. Hence, Equation (3.3) in light of [11, p. 262, Equation (4.99)] yields

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-sV_1(T_1)}|V_1(0)=v) = &\frac{c_1\mathrm{e}^{-sv}}{c_1+\lambda_1(1-\tilde{b}_1(s))-s}\\ &-\int_{t=v}^{\infty} sc_1e^{-c_1t}\int_{u=0}^{t-v} e^{(s-\lambda_1(1-\tilde{b}_1(s)))(t-u-v)}\mathbb{P}(V_1(u+v)=0|V_1(0)=v)\mathrm{d}u\mathrm{d}t.\end{aligned} \tag{3.4}$$

For the calculation of the integrals in the right hand side of Equation (3.4) we use [11, p. 260, Equation (4.92)], for Re $s\geq 0$, $t\geq 0$,

$$\int_{t=0}^{\infty} e^{-st}\mathbb{P}(V_1(t)=0|V_1(0)=v)\mathrm{d}t = \frac{e^{-(s+(1-\mu(s,1))\lambda_1)v}}{s+(1-\mu(s,1))\lambda_1},$$

with $\mu(s,1)$ denoting the LST of the busy period distribution of the M/G/1 queue with arrival rate λ_1 and service time LST $\tilde{b}_1(s)$; $\mu(s,1)$ is the root of $z=\tilde{b}_1\left(s+(1-z)\lambda_1\right)$ with the smallest absolute value, cf. [11, p. 250]). A lengthy but straightforward calculation, that involves interchanging the integrations, yields, for Re $s\geq 0$,

$$\begin{aligned}&\int_{t=v}^{\infty} sc_1e^{-c_1t}\int_{u=0}^{t-v} e^{(s-\lambda_1(1-\tilde{b}_1(s)))(t-u-v)}\mathbb{P}(V_1(u+v)=0|V_1(0)=v)\mathrm{d}u\mathrm{d}t =\\ &\qquad\frac{sc_1}{c_1-s+\lambda_1(1-\tilde{b}_1(s))}\frac{e^{-(c_1+(1-\mu(c_1,1))\lambda_1)v}}{c_1+(1-\mu(c_1,1))\lambda_1}.\end{aligned} \tag{3.5}$$

Combining (3.4) and (3.5) yields

$$\mathbb{E}(e^{-sV_1(T_1)}|V_1(0)=v)=\frac{c_1e^{-sv}}{c_1+\lambda_1(1-\tilde{b}_1(s))-s}-\frac{sc_1}{c_1-s+\lambda_1(1-\tilde{b}_1(s))}\frac{e^{-(c_1+(1-\mu(c_1,1))\lambda_1)v}}{c_1+(1-\mu(c_1,1))\lambda_1}. \tag{3.6}$$

Step 3: Calculation of $\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(0)=v)$.

$$\begin{aligned}\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(0)=v)=&\int_{y=0}^{\infty}\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(T_1)=y)f_{V_1}(V_1(T_1)=y|V_1(0)=v)\mathrm{d}y\\ =&\frac{c_2}{c_2+\lambda_1(1-\tilde{b}_1(s))}\int_{y=0}^{\infty}e^{-sy}f_{V_1}(V_1(T_1)=y|V_1(0)=v)\mathrm{d}y\\ =&\frac{c_2}{c_2+\lambda_1(1-\tilde{b}_1(s))}\left[\frac{e^{-sv}c_1}{c_1+\lambda_1(1-\tilde{b}_1(s))-s}\right.\\ &\left.-\frac{c_1e^{-sv}}{c_1-s+\lambda_1(1-\tilde{b}_1(s))}\frac{e^{-(c_1+(1-\mu(c_1,1))\lambda_1)v}}{c_1+(1-\mu(c_1,1))\lambda_1}\right],\end{aligned} \tag{3.7}$$

where the second equation comes from Equation (3.2) and the third from Equation (3.6).

Step 4: Calculation of $\mathbb{E}(e^{-sV_1(T_1+T_2)})$ in stationarity.
Observe that

$$\begin{aligned}\mathbb{E}(e^{-sV_1(T_1+T_2)})=&\int_{v=0}^{\infty}\mathbb{E}(e^{-sV_1(T_1+T_2)}|V_1(0)=v)f_{V_1(0)}(v)\mathrm{d}v\\ =&\int_{v=0}^{\infty}\left[\frac{c_2}{c_2+\lambda_1(1-\tilde{b}_1(s))}\left[e^{-sv}\frac{c_1}{c_1+\lambda_1(1-\tilde{b}_1(s))-s}\right.\right.\\ &\left.\left.-\ s\frac{c_1}{c_1-s+\lambda_1(1-\tilde{b}_1(s))}\frac{e^{-(c_1+(1-\mu(c_1,1))\lambda_1)v}}{c_1+(1-\mu(c_1,1))\lambda_1}\right]\right]f_{V_1(0)}(v)\mathrm{d}v,\end{aligned} \tag{3.8}$$

with $f_{V_1(0)}(v)$ the probability density function of $V_1(0)$. Now observe that in steady-state $V_1(T_1+T_2)$ has the same distribution as $V_1(0)$. So we can rewrite (3.8) as follows

$$\begin{aligned}\mathbb{E}(e^{-sV_1(T_1+T_2)})=&\int_{v=0}^{\infty}\left[\frac{c_2}{c_2+\lambda_1(1-\tilde{b}_1(s))}\left[e^{-sv}\frac{c_1}{c_1+\lambda_1(1-\tilde{b}_1(s))-s}\right.\right.\\ &\left.\left.-s\frac{c_1}{c_1-s+\lambda_1(1-\tilde{b}_1(s))}\frac{e^{-(c_1+(1-\mu(c_1,1))\lambda_1)v}}{c_1+(1-\mu(c_1,1))\lambda_1}\right]\right]f_{V_1(T_1+T_2)}(v)\mathrm{d}v.\end{aligned}$$

So,

$$\begin{aligned}&\mathbb{E}(e^{-sV_1(T_1+T_2)})\left[\frac{c_2+\lambda_1(1-\tilde{b}_1(s))}{c_2}-\frac{c_1}{c_1+\lambda_1(1-\tilde{b}_1(s))-s}\right]\\ &=-\frac{sc_1}{\left[c_1+\lambda_1(1-\tilde{b}_1(s))-s\right](c_1+(1-\mu(c_1,1))\lambda_1)}\mathbb{E}(e^{-(c_1+(1-\mu(c_1,1))\lambda_1)V_1(T_1+T_2)}).\end{aligned} \tag{3.9}$$

Taking the limit as $s\to 0$ in (3.9) and using L'Hôpital's rule yields

$$\mathbb{E}(e^{-(c_1+(1-\mu(c_1,1))\lambda_1)V_1(T_1+T_2)})=-\frac{[\lambda_1\mathbb{E}(B_1)c_1+\lambda_1\mathbb{E}(B_1)c_2-c_2](c_1+(1-\mu(c_1,1))\lambda_1)}{c_1c_2}.$$

Hence,

$$\mathbb{E}(e^{-sV_1(T_1+T_2)})=\frac{s\left[\lambda_1\mathbb{E}(B_1)(c_1+c_2)-c_2\right]}{\left[c_2+\lambda_1(1-\tilde{b}_1(s))\right]\left[c_1+\lambda_1(1-\tilde{b}_1(s))-s\right]-c_1c_2}. \tag{3.10}$$

Step 5: Calculation of the stationary distribution $\mathbb{E}(e^{-sV_1})$ in steady-state.
Firstly, let us denote by $S = 1$ (respectively by $S = 2$) the event of the server residing in the first (respectively second) queue. Then,

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-sV_1}) &= \mathbb{E}(\mathrm{e}^{-sV_1}|S=1)\mathbb{P}(S=1) + \mathbb{E}(\mathrm{e}^{-sV_1}|S=2)\mathbb{P}(S=2) \\ &= \mathbb{E}(\mathrm{e}^{-sV_1}|S=1)\frac{c_2}{c_1+c_2} + \mathbb{E}(\mathrm{e}^{-sV_1}|S=2)\frac{c_1}{c_1+c_2}. \end{aligned} \tag{3.11}$$

Because of the memoryless property of the exponential distribution it is obvious that

$$\mathbb{E}(\mathrm{e}^{-sV_1}|S=1) = \mathbb{E}(e^{-sV_1(T_1)}),\ \mathbb{E}(\mathrm{e}^{-sV_1}|S=2) = \mathbb{E}(e^{-sV_1(T_1+T_2)}).$$

The latter term is given by (3.10), while the former term is calculated using the same argument as in the derivation of Equation (3.2)

$$\mathbb{E}(e^{-sV_1(T_1+T_2)}) = \mathbb{E}(e^{-sV_1(T_1)})\frac{c_2}{c_2+\lambda_1(1-\tilde{b}_1(s))}. \tag{3.12}$$

Substituting (3.12), for $\mathbb{E}(e^{-sV_1(T_1)})$, and (3.10) in Equation (3.11) yields (3.1), which concludes the proof.

Similarly, we can also calculate the LST of the workload of the second queue. □

Remark 3.1. *It is not difficult to extend the above results to the case that the T_2 periods are non-exponential, see, e.g., [22, 37, 3], and to the case that the arrival process during those periods is a different compound Poisson process than during the T_1 periods, see, e.g., [36] and [19]. One could even allow a more general non-decreasing Lévy process (subordinator) during those T_2 periods. During T_1 periods, one could also allow the input process to be a subordinator. However, we do note that it is considerably more complicated to consider non-exponential T_1 periods, see, [18].*

4 Workload decomposition

In this section, we show that the steady-state workload V_1 (similarly for V_2) can be decomposed into two independent terms, one corresponding to the steady-state workload of the first queue in isolation, i.e., the M/G/1 queue with arrival rate λ_1 and service times B_1, and the second corresponding to the amount of work when the server is not servicing the first queue, due to either an idle period or due to a visit at the second queue.
Using the decomposition of V_1, we determine the mean and the variance of the workload V_1. Furthermore, in the next section, we use the decomposition to obtain various asymptotic (heavy traffic and/or heavy tail) results.

Theorem 4.1. *The steady-state amount of work of the first queue, V_1, is distributed as*

$$V_1 \overset{d}{=} V_{M/G/1} + Y, \tag{4.1}$$

where $V_{M/G/1}$ is the steady-state amount of work in the corresponding M/G/1 queue, and Y is the steady-state amount of work when the server is not servicing at the first queue, with $V_{M/G/1}$ and Y independent.

Proof. The workload decomposition result follows from [7, Theorem 2.1]; it is readily verified that all conditions of that theorem are satisfied. Alternatively, one can also prove (4.1) directly, by focusing on Y and observing that there are two periods in which the server is not working for customers in the first queue: when the server is idle in a visit period, and when the server is visiting the second queue. Accordingly, $\mathbb{E}(e^{-sY})$ can be written as the weighted sum of the LST of the workloads when the server is idle in a visit period, i.e., $\{V_1 = 0|S = 1\} \overset{d}{=} \{V_1(T_1) = 0\}$, and when the server is not at the first queue, i.e., $\{V_1|S = 2\} \overset{d}{=} V_1(T_1 + T_2)$. Thus,

$$\mathbb{E}(e^{-sY}) = \frac{\frac{c_2}{c_1+c_2}\mathbb{P}(V_1(T_1)=0) + \frac{c_1}{c_1+c_2}\mathbb{E}(e^{-sV_1(T_1+T_2)})}{\frac{c_2}{c_1+c_2}\mathbb{P}(V_1(T_1)=0) + \frac{c_1}{c_1+c_2}}. \tag{4.2}$$

Using (3.12) and (3.10), we get

$$\mathbb{P}(V_1(T_1)=0)=\lim_{s\to\infty}\mathbb{E}(e^{-sV_1(T_1)})=1-\rho_1\frac{c_1+c_2}{c_2}. \tag{4.3}$$

Substituting Equation (4.3) in (4.2) gives

$$\mathbb{E}(e^{-sY})=\frac{1}{(1-\rho_1)(c_1+c_2)}\left[c_2-\rho_1(c_1+c_2)+c_1\mathbb{E}(e^{-sV_1(T_1+T_2)})\right]. \tag{4.4}$$

Substituting Equation (3.10) in (4.4) yields

$$\mathbb{E}(e^{-sY})=\frac{c_2-\rho_1(c_1+c_2)}{(1-\rho_1)(c_1+c_2)}\left[1-\frac{sc_1}{(c_2+c_1-s)\lambda_1(1-\tilde{b}(s))+\big(\lambda_1(1-\tilde{b}(s))\big)^2-sc_2}\right]. \tag{4.5}$$

Multiplying the expression above for $\mathbb{E}(e^{-sY})$ with the known LST of the M/G/1 queue (cf. [11, p. 257, Equation (4.90)]) yields Equation (3.1), i.e., the LST of the workload V_1, which confirms that the workload decomposition holds. □

We now use the decomposition result (4.1) to determine the mean and the variance of V_1.

Theorem 4.2. *The expectation of the steady-state workload of the first queue, $\mathbb{E}(V_1)$, is*

$$\mathbb{E}(V_1)=\frac{\rho_1(c_1+c_2)}{c_2-\rho_1(c_1+c_2)}\left[\frac{1}{2}\frac{\mathbb{E}(B_1^2)}{\mathbb{E}(B_1)}+\frac{c_1}{(c_1+c_2)^2}\right], \tag{4.6}$$

and the corresponding variance, $\mathbb{V}\mathrm{ar}(V_1)$, is

$$\mathbb{V}\mathrm{ar}(V_1) = \frac{\rho_1(c_1+c_2)}{c_2-\rho_1(c_1+c_2)}\left[\frac{1}{3}\frac{\mathbb{E}(B_1^3)}{\mathbb{E}(B_1)}+\frac{1}{4}\frac{\rho_1}{1-\rho_1}\frac{\left(\mathbb{E}(B_1^2)\right)^2}{\left(\mathbb{E}(B_1)\right)^2}+\frac{c_1}{(c_1+c_2)^2}\frac{\mathbb{E}(B_1^2)}{\mathbb{E}(B_1)}+\frac{c_1}{(c_1+c_2)^3}\right]. \tag{4.7}$$

Proof. The mean and variance can be obtained by using the decomposition result (4.1). For this purpose, we can separately calculate the mean and the variance of the M/G/1 queue, cf. [11, p. 256], as well as the mean and the variance corresponding to the random variable Y. For the latter we use Equation (4.5) (after dividing its numerator and denominator by s). Combining these results yields Equations (4.6) and (4.7). □

Remark 4.1. *Equation (4.6) and Equation (4.7) for $c_2\to\infty$ (or equivalently for $c_1\to 0$) yield the corresponding expressions for the mean and the variance of the M/G/1 queue, cf. [11, p. 256].*

5 Heavy traffic and heavy tail asymptotics

In this section, we study the heavy traffic behavior of the workload V_1, both for the case of light tailed and of heavy tailed distribution of the service time B_1. We also discuss the tail behavior of the workload in the case of regularly varying (and subexponential) service time distributions.

We first consider the behavior of V_1 in heavy traffic, i.e., when $\rho_1\uparrow\frac{c_2}{c_1+c_2}$. In Theorem 4.1 we have shown that V_1 can be written as the sum of the independent random variables $V_{\text{M/G/1}}$ and Y. Since most of the results related to the M/G/1 queue are already known, we take a closer look at $\mathbb{E}(e^{-sY})$, assuming for now that the first three moments of B_1 are finite. Substituting $\tilde{b}(s)=1-s\mathbb{E}(B_1)+\frac{s^2}{2}\mathbb{E}(B_1^2)-\frac{s^3}{3!}\mathbb{E}(B_1^3)+o(s^3)$ in (4.5) and rearranging the terms yields

$$\mathbb{E}(e^{-sY})=\frac{A_0}{1-\rho_1}\left[1-\frac{c_1}{c_1+c_2}\left(\frac{1}{A_0+sA_1-\frac{s^2}{2}A_2+o(s^2)}\right)\right], \tag{5.1}$$

with

$$A_0 = \frac{c_2}{c_1+c_2} - \rho_1,\ A_1 = \frac{\rho_1}{c_1+c_2}\left(1-\rho_1+\frac{c_1+c_2}{2}\frac{\mathbb{E}(B_1^2)}{\mathbb{E}(B_1)}\right), \tag{5.2}$$

$$A_2 = \frac{\rho_1^2}{c_1+c_2}\left(\frac{1-2\rho_1}{\rho_1}\frac{\mathbb{E}(B_1^2)}{\mathbb{E}(B_1)}+\frac{c_1+c_2}{3\rho_1}\frac{\mathbb{E}(B_1^3)}{\mathbb{E}(B_1)}\right). \tag{5.3}$$

Equation (5.1) will play a very important role in the proofs of the theorems of this section regarding the asymptotic behavior of the workload V_1.

Theorem 5.1. *Assume that $\mathbb{E}(B_1^2) < \infty$. For $\rho_1 \uparrow \frac{c_2}{c_1+c_2}$,*

$$\left(\frac{c_2}{c_1+c_2} - \rho_1\right) V_1 \xrightarrow{d} Z, \tag{5.4}$$

with Z an exponentially distributed random variable with mean $\frac{c_1 c_2}{(c_1+c_2)^3} + \frac{c_2}{c_1+c_2}\frac{1}{2}\frac{\mathbb{E}(B_1^2)}{\mathbb{E}(B_1)}$.

Proof. To obtain the heavy traffic limit of V_1 one can use the workload decomposition. Theorem 4.1 implies that

$$\mathbb{E}(e^{-sV_1}) = \mathbb{E}(e^{-sV_{\mathrm{M/G/1}}})\mathbb{E}(e^{-sY}). \tag{5.5}$$

Replacing s by $sA_0 = s(\frac{c_2}{c_1+c_2} - \rho_1)$, cf. (5.2), in the above equation and taking the limit $\rho_1 \uparrow \frac{c_2}{c_1+c_2}$ yields

$$\lim_{\rho_1 \uparrow \frac{c_2}{c_1+c_2}} \mathbb{E}\left(e^{-s\left(\frac{c_2}{c_1+c_2}-\rho_1\right)V_1}\right) = \lim_{\rho_1 \uparrow \frac{c_2}{c_1+c_2}} \mathbb{E}\left(e^{-s\left(\frac{c_2}{c_1+c_2}-\rho_1\right)V_{\mathrm{M/G/1}}}\right)\mathbb{E}\left(e^{-s\left(\frac{c_2}{c_1+c_2}-\rho_1\right)Y}\right). \tag{5.6}$$

The first term in the right hand side obviously tends to one for $\rho_1 \uparrow \frac{c_2}{c_1+c_2}$, as the corresponding M/G/1 queue is in heavy traffic only when $\rho_1 \uparrow 1$. In order to calculate the limit for the second term in (5.6), we replace s by $sA_0 = s(\frac{c_2}{c_1+c_2} - \rho_1)$, cf. (5.2), in (5.1), which yields

$$\mathbb{E}(e^{-s(\frac{c_2}{c_1+c_2}-\rho_1)Y}) = \frac{1}{1-\rho_1}\left[A_0 + \frac{c_1}{c_1+c_2}\left(\frac{1}{1+sA_1-\frac{s^2}{2}A_0A_2+o\left(s^2A_0\right)}\right)\right]. \tag{5.7}$$

Taking the limit $\rho_1 \uparrow \frac{c_2}{c_1+c_2}$ in (5.7) yields

$$\lim_{\rho_1 \uparrow \frac{c_2}{c_1+c_2}} \mathbb{E}(e^{-s\left(\frac{c_2}{c_1+c_2}-\rho_1\right)Y}) = \frac{1}{1+sA_1}, \tag{5.8}$$

with A_1 given in (5.2). From (5.8), (5.2), and (5.6) the statement of the theorem follows by noticing that the right hand side of (5.8) corresponds to the LST of an exponentially distributed random variable with mean A_1. □

Remark 5.1. *Letting $c_2 \to \infty$, Theorem 5.1 indicates that the heavy traffic result reduces to that of an ordinary M/G/1 queue.*

We now turn our focus to the asymptotic workload analysis in the case of a heavy tailed service time distribution. First, we introduce the definition of a regularly varying random variable/distribution.

Definition 5.1. *The distribution function of a random variable B_1 on $[0,\infty)$ is called regularly varying of index $-\nu$, with $\nu \in \mathbb{R}$, if*

$$\mathbb{P}(B_1 > x) \sim L(x)x^{-\nu},\ x \uparrow \infty, \tag{5.9}$$

with $L(x)$ a slowly varying function at infinity, i.e., $\lim_{x\to\infty}\frac{L(\alpha x)}{L(x)} = 1$, for all $\alpha > 1$.

Theorem 5.2. *If the service time distribution of the random variable B_1 is regularly varying of index $-\nu$, with $\nu \in (1,2)$, then the workload of the first queue under the stability condition (2.1) is regularly varying at infinity of index $1-\nu$. More precisely,*

$$\mathbb{P}(V_1 > x) \sim \frac{\rho_1}{\frac{c_2}{c_1+c_2} - \rho_1}\frac{1}{\mathbb{E}(B_1)(\nu-1)}x^{1-\nu}L(x), \qquad x \uparrow \infty. \tag{5.10}$$

Proof. To prove that V_1 is regularly varying at infinity, one can again use the decomposition property of the workload V_1. From Theorem 4.1, we get

$$\mathbb{P}(V_1 > x) = \mathbb{P}(V_{\text{M/G/1}} + Y > x). \tag{5.11}$$

In the M/G/1 queue it follows from [10] that $\mathbb{P}(V_{\text{M/G/1}} > x)$ is regularly varying of index $1-\nu$ at infinity if and only if the tail of the service time distribution $\mathbb{P}(B_1 > x)$ is regularly varying of index $-\nu$ at infinity, and one has

$$\mathbb{P}(V_{\text{M/G/1}} > x) \sim \frac{\rho_1}{\rho_1 - 1} \frac{1}{\mathbb{E}(B_1)(1-\nu)} x^{1-\nu} L(x), \qquad x \uparrow \infty. \tag{5.12}$$

Now we have to compute $\mathbb{P}(Y > x)$ for $x \uparrow \infty$. Our main tool is the Tauberian theorem of [4, Theorem 8.1.6], which relates the behavior of a regularly varying function at infinity and the behavior of its LST near 0. Applying this theorem to (5.9) gives

$$\tilde{b}_1(s) - 1 + s\mathbb{E}(B_1) \sim -\Gamma(1-\nu) s^{\nu} L\left(\frac{1}{s}\right), \qquad s \downarrow 0,$$

and hence

$$\frac{\lambda_1 \left(1 - \tilde{b}_1(s)\right)}{s} = \rho_1 \left(1 + \frac{\Gamma(1-\nu)}{\mathbb{E}(B_1)} s^{\nu-1} L\left(\frac{1}{s}\right)\right), \qquad s \downarrow 0. \tag{5.13}$$

Substituting Equation (5.13) in (4.5) yields, for $s \downarrow 0$:

$$\begin{aligned}
\mathbb{E}(e^{-sY}) &= \frac{c_2 - \rho_1(c_1+c_2)}{(1-\rho_1)(c_1+c_2)} \left[1 - \frac{c_1}{(c_1+c_2)\rho_1\left(1 + \frac{\Gamma(1-\nu)}{\mathbb{E}(B_1)} s^{\nu-1} L\left(\frac{1}{s}\right)\right) - c_2 + O(s)}\right] \\
&= \frac{c_2 - \rho_1(c_1+c_2)}{(1-\rho_1)(c_1+c_2)} \left[1 + \frac{c_1}{(c_2 - \rho_1(c_1+c_2))\left(1 - \frac{\rho_1(c_1+c_2)}{c_2-\rho_1(c_1+c_2)} \frac{\Gamma(1-\nu)}{\mathbb{E}(B_1)} s^{\nu-1} L\left(\frac{1}{s}\right)\right) + O(s)}\right] \\
&= \frac{c_2 - \rho_1(c_1+c_2)}{(1-\rho_1)(c_1+c_2)} \\
&\quad + \frac{c_1}{(1-\rho_1)(c_1+c_2)} \left(1 + \frac{\rho_1(c_1+c_2)}{c_2 - \rho_1(c_1+c_2)} \frac{\Gamma(1-\nu)}{\mathbb{E}(B_1)} s^{\nu-1} L\left(\frac{1}{s}\right) + O(s)\right).
\end{aligned}$$

Simplifying, we get

$$\mathbb{E}(e^{-sY}) - 1 = \frac{\rho_1 c_1}{(1-\rho_1)(c_2 - \rho_1(c_1+c_2))} \frac{\Gamma(1-\nu)}{\mathbb{E}(B_1)} s^{\nu-1} L\left(\frac{1}{s}\right), \qquad s \downarrow 0.$$

Applying the Tauberian theorem of [4, Theorem 8.1.6] once again, now in the reverse direction, yields

$$\begin{aligned}
\mathbb{P}(Y > x) &\sim -\frac{1}{\Gamma(2-\nu)} \frac{\rho_1 c_1}{(1-\rho_1)(c_2 - \rho_1(c_1+c_2))} \frac{\Gamma(1-\nu)}{\mathbb{E}(B_1)} x^{1-\nu} L(x) \\
&= \frac{\rho_1 c_1}{(1-\rho_1)(c_2 - \rho_1(c_1+c_2))} \frac{1}{\mathbb{E}(B_1)(\nu-1)} x^{1-\nu} L(x), \qquad x \uparrow \infty.
\end{aligned} \tag{5.14}$$

From Equation (5.12) and (5.14), we see that both $V_{\text{M/G/1}}$ and Y are regularly varying random variables of index $1-\nu$. Using the workload decomposition property (4.1) and a well known result regarding the tail behavior of the sum of two independent regularly varying random variables of the same index, see [30], yields

$$\mathbb{P}(V_1 > x) \sim (C_1 + C_2) x^{1-\nu} L(x), \qquad x \uparrow \infty, \tag{5.15}$$

with C_1 and C_2 the coefficients of the tail $x^{1-\nu}$ for $V_{\text{M/G/1}}$ and Y in (5.12) and (5.14), respectively. Substituting the coefficients from (5.12) and (5.14) concludes the proof of the theorem. □

Remark 5.2. *Letting $c_2 \to \infty$ in Equation (5.15) yields*

$$\mathbb{P}(V_1 > x) = \frac{\rho_1}{1-\rho_1} \frac{1}{\mathbb{E}(B_1)(\nu-1)} x^{1-\nu} L(x), \qquad x \uparrow \infty, \tag{5.16}$$

which is, as expected, the result for an ordinary M/G/1 queue.

Remark 5.3. *Theorem 5.2 is closely related to [9, Theorem 4.1] for a single server queue with alternating high and low service speeds. In [9] both the service time distribution* and *the distribution of the periods of low service speed are regularly varying. If the latter tail is less heavy than the tail of the service time distribution, then our formula (5.10) displays exactly the same tail behavior as [9, Formula (4.1)].*

We briefly discuss how Theorem 5.2 can be generalized to the case of subexponential (residual) service times.

Definition 5.2. *A distribution function $\mathbb{P}(B_1 \leq x)$, $x \geq 0$, is called subexponential if*

$$\mathbb{P}(B_{11} + \cdots + B_{1n} > x) \sim n\mathbb{P}(B_{11} > x),\ x \uparrow \infty,$$

for any $n \geq 2$, with $B_{11}, \ldots, B_{1n}$ independent and identical copies of B_1.

It can be shown that a similar result as in Theorem 5.2 holds for subexponential distributions. Specifically, if the distribution of the residual service time requirement, say B_1^r, is subexponential, then V_1 is also subexponential and

$$\mathbb{P}(V_1 > x) \sim \frac{\rho_1}{\frac{c_2}{c_1+c_2} - \rho_1}\mathbb{P}(B_1^r > x),\ x \uparrow \infty. \tag{5.17}$$

Heuristic proof. The asymptotic relation in (5.17) can be proved formally using sample-path techniques along the following lines. We assume the system is in stationarity and focus on the workload at time $t = 0$. If the workload level V_1 at this time is very large, then that is most likely due to the prior arrival of a customer with a large service requirement B_1, at some time $t = -y$. We can observe that from time $t = -y$ onward, the workload decreases nearly linearly with rate $\frac{c_2}{c_1+c_2} - \rho_1$. So in order for the workload at time $t = 0$ to exceed the level x, the service requirement B_1 must be larger than $x + y\left(\frac{c_2}{c_1+c_2} - \rho_1\right)$. Since customers arrive according to a Poisson process with rate λ_1, the distribution of the workload V_1 for large x can be computed as

$$\mathbb{P}(V_1 > x) \sim \int_{y=0}^{\infty} \mathbb{P}\left(B_1 > x + y\left(\frac{c_2}{c_1+c_2} - \rho_1\right)\right)\lambda_1 \mathrm{d}y. \tag{5.18}$$

A change of variable $z := x + y\left(\frac{c_2}{c_1+c_2} - \rho_1\right)$ in Equation (5.18) yields

$$\begin{aligned}\mathbb{P}(V_1 > x) &\sim \frac{\lambda_1}{\frac{c_2}{c_1+c_2} - \rho_1}\int_{z=x}^{\infty} \mathbb{P}(B_1 > z)\,\mathrm{d}z \\ &= \frac{\lambda_1 \mathbb{E}(B_1)}{\frac{c_2}{c_1+c_2} - \rho_1}\int_{z=x}^{\infty} \frac{\mathbb{P}(B_1 > z)}{\mathbb{E}(B_1)}\mathrm{d}z = \frac{\rho_1}{\frac{c_2}{c_1+c_2} - \rho_1}\mathbb{P}(B_1^r > x),\end{aligned} \tag{5.19}$$

which leads to Relation (5.17).
This proof can be made rigorous by providing lower and upper bounds for $\mathbb{P}(V_1 > x)$ that in the limit coincide. The lower bound is easily obtained by using the law of large numbers. The upper bound is more difficult; one has to give a formal version of the statement *"exceedance of a high level x happens as a consequence of a single big jump"*, and one has to show that other exceedance scenarios (like two rather big jumps) do not contribute to the asymptotics of the exceedance probability. We refer to [41, Section 2.4] for a detailed exposition of this technique.

Remark 5.4. *Note that, indeed, Relation (5.17) contains the result of Theorem 5.2 as a special case, since B_1 being regularly varying at infinity of index $-\nu$, with $\nu \in (1,2)$, has a subexponential distribution. In this regularly varying case, we have*

$$\mathbb{P}(B_1^r > x) = \int_{z=x}^{\infty} \frac{\mathbb{P}(B_1 > z)}{\mathbb{E}(B_1)}\mathrm{d}z \sim \frac{1}{\mathbb{E}(B_1)}\int_{z=x}^{\infty} z^{-\nu}L(z)\mathrm{d}z \sim \frac{1}{\mathbb{E}(B_1)(\nu-1)}x^{1-\nu}L(x),\ x \uparrow \infty, \tag{5.20}$$

where the last step in (5.20) follows from [41, p. 26, Lemma 2.1.7]. Combining (5.19) and (5.20), we obtain Theorem 5.2.

Now we have all necessary ingredients to state and prove a heavy traffic limit theorem for V_1 in the heavy tailed case. To do this analysis, we first scale V_1 by the coefficient of contraction $\Delta(\rho_1)$. Similarly to [8, p. 188, Equation (4.24)], we define the coefficient of contraction $\Delta(\rho_1)$ as the unique root of the following equation in x

$$x^{\nu-1}L\left(\frac{1}{x}\right)=\frac{\frac{c_2}{c_1+c_2}-\rho_1}{\rho_1},\ x>0, \tag{5.21}$$

such that $\Delta(\rho_1)\downarrow 0$ for $\rho_1\uparrow\frac{c_2}{c_1+c_2}$.

Theorem 5.3. *If the service time distribution of the random variable B_1 is regularly varying of index $-\nu$, with $\nu\in(1,2)$, then the heavy traffic limiting distribution of workload V_1 of the first queue in the heavy tailed case is given by the Mittag-Leffler distribution*

$$\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\mathbb{E}\left(e^{-s\Delta(\rho_1)V_1}\right)=\frac{1}{1+(\mathbb{E}(B_1))^{\nu-1}s^{\nu-1}}. \tag{5.22}$$

Proof. We can obtain the heavy traffic limit in the heavy tailed case by using the workload decomposition property (4.1) and its LST version (5.5). The heavy traffic limit can be computed by replacing s by $s\Delta(\rho_1)$ in Equation (5.5) and taking the limit $\rho_1\uparrow\frac{c_2}{c_1+c_2}$, which yields

$$\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\mathbb{E}\left(e^{-s\Delta(\rho_1)V_1}\right)=\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\mathbb{E}\left(e^{-s\Delta(\rho_1)V_{\text{M/G/1}}}\right)\mathbb{E}\left(e^{-s\Delta(\rho_1)Y}\right). \tag{5.23}$$

Just as in the light tailed case (cf. Theorem 5.1), the contribution of $V_{\text{M/G/1}}$ becomes negligible compared to the contribution of Y. To calculate the limit for the second factor in (5.23), we use (4.5)

$$\mathbb{E}\left(e^{-s\Delta(\rho_1)Y}\right)=\frac{\frac{c_2}{c_1+c_2}-\rho_1}{1-\rho_1}\left[1-\frac{c_1}{c_1+c_2}\frac{1}{\frac{f(s\Delta(\rho_1))}{s\Delta(\rho_1)}-\frac{f(s\Delta(\rho_1))}{c_1+c_2}+\frac{s\Delta(\rho_1)}{c_1+c_2}\left(\frac{f(s\Delta(\rho_1))}{s\Delta(\rho_1)}\right)^2-\frac{c_2}{c_1+c_2}}\right], \tag{5.24}$$

with $f(s\Delta(\rho_1))=\frac{\rho_1(1-\tilde{b}_1(s\Delta(\rho_1)))}{\mathbb{E}(B_1)}$. Taking the limit $\rho_1\uparrow\frac{c_2}{c_1+c_2}$ in (5.24) yields

$$\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\mathbb{E}\left(e^{-s\Delta(\rho_1)Y}\right)=-\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\frac{c_1\left(\frac{c_2}{c_1+c_2}-\rho_1\right)}{(c_1+c_2)(1-\rho_1)}\frac{1}{\frac{f(s\Delta(\rho_1))}{s\Delta(\rho_1)}-\frac{c_2}{c_1+c_2}}, \tag{5.25}$$

since $\frac{s\Delta(\rho_1)}{c_1+c_2}\left(\frac{f(s\Delta(\rho_1))}{s\Delta(\rho_1)}\right)^2\to 0$, $f(s\Delta(\rho_1))\to 0$ and $\Delta(\rho_1)\to 0$ when $\rho_1\uparrow\frac{c_2}{c_1+c_2}$. After rearranging the terms of (5.25) we get

$$\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\mathbb{E}\left(e^{-s\Delta(\rho_1)Y}\right)=-\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\frac{c_1}{(c_1+c_2)(1-\rho_1)}\frac{1}{\frac{1}{\frac{c_2}{c_1+c_2}-\rho_1}\left[\frac{f(s\Delta(\rho_1))}{s\Delta(\rho_1)}-\frac{c_2}{c_1+c_2}\right]}. \tag{5.26}$$

Since B_1 is regularly varying, we get by using [8, Lemma 5.1 (iv)],

$$\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\frac{1}{\frac{c_2}{c_1+c_2}-\rho_1}\left[\frac{f(s\Delta(\rho_1))}{s\Delta(\rho_1)}-\frac{c_2}{c_1+c_2}\right]=-\lim_{\rho_1\uparrow\frac{c_2}{c_1+c_2}}\left(1+\frac{\rho_1}{\frac{c_2}{c_1+c_2}-\rho_1}\left[1-\frac{1-\tilde{b}_1(s\Delta(\rho_1))}{s\mathbb{E}(B_1)\Delta(\rho_1)}\right]\right). \tag{5.27}$$

Using [8, p. 188, Equation (4.22)], we know that

$$1-\frac{1-\tilde{b}_1(s\Delta(\rho_1))}{s\mathbb{E}(B_1)\Delta(\rho_1)}\sim(\mathbb{E}(B_1)s\Delta(\rho_1))^{\nu-1}L\left(\frac{1}{s\mathbb{E}(B_1)\Delta(\rho_1)}\right),\ s\downarrow 0. \tag{5.28}$$

From the definition of the coefficient of contraction $\Delta(\rho_1)$ as the unique root of Equation (5.21) such that $\Delta(\rho_1)\downarrow 0$ for $\rho_1\uparrow\frac{c_2}{c_1+c_2}$, we have

$$(\Delta(\rho_1))^{\nu-1}L\left(\frac{1}{\Delta(\rho_1)}\right)=\frac{\frac{c_2}{c_1+c_2}-\rho_1}{\rho_1}. \tag{5.29}$$

Furthermore, from the definition of a slowly varying function $L(\cdot)$, we get $\frac{L\left(\frac{1}{s\mathbb{E}(B_1)\Delta(\rho_1)}\right)}{L\left(\frac{1}{\Delta(\rho_1)}\right)} \to 1$, when $\Delta(\rho_1) \downarrow 0$. Now by combining (5.26) - (5.29) we get

$$\lim_{\rho_1 \uparrow \frac{c_2}{c_1+c_2}} \mathbb{E}\left(e^{-s\Delta(\rho_1)Y}\right) = \frac{1}{1+(\mathbb{E}(B_1))^{\nu-1}s^{\nu-1}}. \tag{5.30}$$

Substituting (5.30) in Equation (5.23) concludes the proof of the theorem. □

Remark 5.5. *In [8] a class of service time distributions is considered that is slightly larger than the class of regularly varying distributions. Theorem 5.3 can be seen to hold under these conditions as well.*

6 Joint workload distribution

So far we have focused on the marginal workload distribution at the first queue. A much harder problem is to determine the steady-state joint workload distribution. In this section, we begin the exploration of this problem, outlining a possible approach as well as the mathematical complications arising.

Let $\tilde{v}(s_1,s_2) := \mathbb{E}(\mathrm{e}^{-s_1V_1(T_1+T_2)-s_2V_2(T_1+T_2)})$ denote the steady-state joint workload LST at endings of visit periods at the second queue. Reiterating Steps 1 - 4 of Section 3, but now taking both workloads into account, leads after lengthy calculations to the following functional equation

$$\tilde{v}(s_2,s_1) = \frac{c_1}{c_1-s_1+\lambda_1(1-\tilde{b}_1(s_1))+\lambda_2(1-\tilde{b}_2(s_2))}[\tilde{v}(s_1,s_2) - \frac{s_1}{\omega_1(s_2)}\tilde{v}(\omega_1(s_2),s_2)], \quad \text{Re } s_1,\ \text{Re } s_2 \geq 0, \tag{6.1}$$

with $\omega_1(s_2) := c_1+\lambda_2(1-\tilde{b}_2(s_2))+\lambda_1(1-\mu(\zeta,1))$; as before, $\mu(s,1)$ is the busy period LST of the M/G/1 queue in isolation corresponding to the first queue.

Let us now restrict ourselves to the fully symmetric case $c_1 = c_2 = c$, $\lambda_1 = \lambda_2 = \lambda$, $\tilde{b}_1(s) = \tilde{b}_2(s) = \tilde{b}(s)$. Formula (6.1) then becomes

$$\tilde{v}(s_2,s_1) = \frac{c}{c-s_1+\lambda(1-\tilde{b}(s_1))+\lambda(1-\tilde{b}(s_2))}[\tilde{v}(s_1,s_2) - \frac{s_1}{\omega_1(s_2)}\tilde{v}(\omega_1(s_2),s_2)]. \tag{6.2}$$

Taking $s_1 = s_2$ in (6.2) allows us to express $\tilde{v}(\omega_1(s_2),s_2)$ in terms of $\tilde{v}(s_2,s_2)$, thus reducing (6.2) to

$$\tilde{v}(s_2,s_1) = \frac{c}{c-s_1+\lambda(1-\tilde{b}(s_1))+\lambda(1-\tilde{b}(s_2))}[\tilde{v}(s_1,s_2) - \frac{s_1}{s_2}\frac{s_2-2\lambda(1-\tilde{b}(s_2))}{c}\tilde{v}(s_2,s_2)]. \tag{6.3}$$

Interchanging all indices, one obtains a mirrored equation of (6.3), and the two equations combined yield

$$\begin{aligned} K(s_1,s_2)\tilde{v}(s_1,s_2) = &\frac{s_2}{s_1}\left(s_1-2\lambda(1-\tilde{b}(s_1))\right)\left(c-s_1+\lambda(1-\tilde{b}(s_1))+\lambda(1-\tilde{b}(s_2))\right)\tilde{v}(s_1,s_1) \\ &+\frac{s_1}{s_2}c\left(s_2-2\lambda(1-\tilde{b}(s_2))\right)\tilde{v}(s_2,s_2), \ \text{Re } s_1,\ \text{Re } s_2 \geq 0, \end{aligned} \tag{6.4}$$

with $K(s_1,s_2) = c^2 - \left(c-s_1+\lambda(1-\tilde{b}(s_1))+\lambda(1-\tilde{b}(s_2))\right)\left(c-s_2+\lambda(1-\tilde{b}(s_1))+\lambda(1-\tilde{b}(s_2))\right)$. This is a so-called boundary value problem equation. Equations of this type have been studied in the monograph [12]. There an approach is outlined that, for the present problem, amounts to the following global steps:

Step 1: Consider the zeros of the *kernel* equation $K(s_1,s_2)$, that have Re s_1, Re $s_2 \geq 0$. For such pairs (s_1,s_2), $\tilde{v}(s_1,s_2)$ is analytic, and hence, for those pairs, the right hand side of (6.4) is equal to zero.

Step 2: For the pairs (s_1,s_2) satisfying Step 1, one needs to translate the fact that the right hand side of Equation (6.4) is zero into a Riemann or Riemann-Hilbert boundary value problem. The solution of such a problem yields $\tilde{v}(s_1,s_1)$ and $\tilde{v}(s_2,s_2)$. Then $\tilde{v}(s_1,s_2)$ follows via (6.4).

Unfortunately, the above steps do not constitute a simple, straightforward recipe. For example, several choices of zero pairs are possible in the present problem, and it is not a priori clear what is the best choice. A natural choice, due to the symmetry of the underlying problem, seems to be to restrict oneself to complex conjugate points, i.e., choose $(s_1, s_2) = (z, \bar{z})$. The *kernel* then becomes

$$K(z,\bar{z}) = c^2 - \left(c - z + 2\lambda \mathrm{Re}(1 - \tilde{b}(z))\right)\left(c - \bar{z} + 2\lambda \mathrm{Re}(1 - \tilde{b}(z))\right).$$

Taking

$$c - z + 2\lambda \mathrm{Re}(1 - \tilde{b}(z)) = c\mathrm{e}^{i\theta}, \quad c - \bar{z} + 2\lambda \mathrm{Re}(1 - \tilde{b}(z)) = c\mathrm{e}^{-i\theta}, \ \theta \in [0, 2\pi], \tag{6.5}$$

indeed yields $K(z,\bar{z}) = K(z(\theta), \bar{z}(\theta)) = 0$, for all $\theta \in [0, 2\pi]$, while it is readily checked that for each such θ there is a unique $z(\theta)$ with Re $z(\theta) \geq 0$.

Turning to Step 2, one sees that the $z(\theta)$ satisfying (6.5) describe a closed contour, say L, in the right half plane, for $\theta : 0 \to 2\pi$, while the fact that the right hand side of (6.4) is zero for all these $(s_1, s_2) = (z(\theta), \bar{z}(\theta))$ translates into the following relation

$$\mathrm{Re}\left[\frac{z}{\bar{z}}\left(z - 2\lambda(1 - \tilde{b}(z))\right)\tilde{v}(z,z)\mathrm{e}^{\frac{1}{2}i\theta}\right] = 0, \ z \in L, \tag{6.6}$$

with $\tilde{v}(z,z)$ and $\tilde{b}(z)$ analytic inside L. If the whole expression inside the square brackets of (6.6) would have been analytic inside L, or would have been analytic except for a pole, then we would have obtained a Riemann-Hilbert problem on contour L, see, e.g., [21, Chapters II and IV] or [12, Section I.3]. The solution of such a problem is known when L is the unit circle. For other closed contours, one needs a conformal mapping of that contour onto the unit circle; several procedures are available for obtaining such conformal mappings. Of course $\bar{z}$ is *not* analytic, so we have not yet arrived at a standard Riemann-Hilbert boundary value problem. Just like with the Wiener-Hopf technique in the related Wiener-Hopf boundary value problem, there might be a way around this by applying a suitable (Wiener-Hopf) factorization; this is a path we would like to explore in future research.

Remark 6.1. *There are several open problems emerging at this point. When we manage to solve the present symmetric problem, we are still faced with the more general asymmetric two-queue problem. Subsequently, one could turn to the joint* queue length *distribution. However, a complication there is that a switch between queues might occur during a service time, forcing one to keep track of the length of the residual service time. From that perspective, workload seems to be an easier quantity than queue length.*

In the next section, we turn our focus to the joint queue length distribution, but restricting ourselves to exponential service time distributions, so we do not need to keep track of the residual service time. Instead of pursuing a boundary value approach, we explore a perturbation approach, which allows us to derive an analytic expansion for the joint queue length distribution.

7 Joint queue length distribution

In this section, we turn our attention to the steady-state joint queue length distribution, restricting ourselves to *exponential service requirement distributions* in both queues, with rates $\mu_i = 1/\mathbb{E}(B_i)$, $i = 1, 2$, respectively. Under this assumption, we do not need to keep track of the residual service times, which simplifies the analysis. However, a direct analytic derivation of the joint queue length distribution (or its PGF) turns out to be as challenging as the analysis presented in Section 6. To address this issue, in this section, we explore the use of parametric perturbation for the derivation of the joint queue length distribution. In what follows, we use the framework developed in Altman et al. [2]; we perturb the service and arrival rates by a common parameter, denoted by $\varepsilon \geq 0$, i.e., the perturbed service rate of the customers in queue i is $\varepsilon\mu_i$, $i = 1, 2$, and arrivals occur according to two independent Poisson processes with perturbed rates $\varepsilon\lambda_i$, $i = 1, 2$. The parameters that are not perturbed are $c_i, i = 1, 2$ i.e., the rates of the exponentially distributed durations that the server spends in each queue. Note that the stability condition (2.1) is not affected by this scaling.

The perturbed process is a continuous time Markov chain defined on the state space

$$\mathscr{S} = \{(n_1,n_2,k),\ n_1,n_2 \in \mathbb{N},\ k \in \{1,2\}\},$$

in which n_i denotes the queue length in queue i, $i = 1,2$, and the third element in the state space description reports the queue in which the server is active. Furthermore, let $\boldsymbol{G}(\varepsilon)$ denote the generator of the perturbed Markov process. We decompose this perturbed generator into the unperturbed generator $\boldsymbol{G}^{(0)}$ and the perturbation matrix $\boldsymbol{G}^{(1)}$,

$$\boldsymbol{G}(\varepsilon) = \boldsymbol{G}^{(0)} + \varepsilon \boldsymbol{G}^{(1)}, \tag{7.1}$$

so as to investigate the dependence of the stationary joint queue length distribution on the parameter ε. The unperturbed generator $\boldsymbol{G}^{(0)}$ corresponds to the states depicting a change of the state of the server from one queue to the other; it is given by

$$\boldsymbol{G}^{(0)} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{0}_{2\times 2} & \cdots \\ \boldsymbol{0}_{2\times 2} & \boldsymbol{C} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{7.2}$$

with $\boldsymbol{C} = \begin{bmatrix} -c_1 & c_1 \\ c_2 & -c_2 \end{bmatrix}$, and $\boldsymbol{0}_{2\times 2}$ a 2×2 matrix of zeros. Throughout the remainder of the paper we use this notation with subscripts to indicate the dimension when needed. When the dimension is clear from the context, the index is omitted; note that the dimension can be infinite.

The perturbation matrix $\boldsymbol{G}^{(1)}$ is defined in terms of its elements, with $n_1 \geq 0$, $n_2 \geq 0$, $k = 1,2$,

$$\begin{aligned} \boldsymbol{G}^{(1)}_{(n_1,n_2,k),(n_1+1,n_2,k)} &= \lambda_1,\ \boldsymbol{G}^{(1)}_{(n_1,n_2,k),(n_1,n_2+1,k)} = \lambda_2, \\ \boldsymbol{G}^{(1)}_{(n_1+1,n_2,1),(n_1,n_2,1)} &= \mu_1,\ \boldsymbol{G}^{(1)}_{(n_1,n_2+1,2),(n_1,n_2,2)} = \mu_2, \\ \boldsymbol{G}^{(1)}_{(n_1,n_2,k),(n_1,n_2,k)} &= -\left(\lambda_1 + \lambda_2 + \mu_k \mathbb{1}_{\{n_k \geq 1\}}\right), \end{aligned} \tag{7.3}$$

with $\mathbb{1}_{\{n_k \geq 1\}}$ an indicator function taking value 1, if $n_k \geq 1$, and 0, otherwise.

In order to implement the framework of Altman et al. [2], it is convenient to first define the transition probability matrix $\boldsymbol{P}(\varepsilon) = \boldsymbol{I} + \Delta \boldsymbol{G}(\varepsilon)$ of the corresponding (uniformized) discrete time perturbed Markov chain ($\boldsymbol{I}$ being the identity matrix). In order to simplify notation, in what follows, we assume without loss of generality that

$$\lambda_1 + \lambda_2 + \mu_1 + c_1 \leq 1 \text{ and } \lambda_1 + \lambda_2 + \mu_2 + c_2 \leq 1. \tag{7.4}$$

Note that indeed, this assumption simply entails a scaling of time. Still, it allows us to take $\Delta = 1$ and it ensures that

$$\boldsymbol{P}(\varepsilon) = \boldsymbol{I} + \boldsymbol{G}(\varepsilon), \tag{7.5}$$

is a probability matrix for all $\varepsilon \in [0,1]$, which is convenient. We remind the reader that our ultimate goal is to find (or approximate) the stationary measure belonging to $\boldsymbol{G}(1)$ (and, equivalently, of the discrete time counter part $\boldsymbol{P}(1)$). In order to achieve that, we first establish the analyticity of the stationary distribution for ε in a punctured neighborhood of 0, cf. Theorem 7.1 below. We emphasize that it is not guaranteed that the stationary distribution will be analytic up to $\varepsilon = 1$. The analysis in [2] gives a lower bound for the radius of convergence, which in general turns out to be rather conservative.

Note that the perturbed transition probability matrix $\boldsymbol{P}(\varepsilon)$ can also be decomposed into the unperturbed probability matrix $\boldsymbol{P}^{(0)}$ and the perturbation matrix $\boldsymbol{G}^{(1)}$, with $\boldsymbol{P}^{(0)} = \boldsymbol{I} + \boldsymbol{G}^{(0)}$, i.e.,

$$\boldsymbol{P}^{(0)} = \begin{bmatrix} \boldsymbol{I}_{2\times 2} + \boldsymbol{C} & \boldsymbol{0}_{2\times 2} & \cdots \\ \boldsymbol{0}_{2\times 2} & \boldsymbol{I}_{2\times 2} + \boldsymbol{C} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}. \tag{7.6}$$

It is evident that the unperturbed process consists of several ergodic classes, making our setting fit the singular perturbation approach in [2].

7.1 Singular perturbation analysis: outline

Following the analysis performed in [2], we now formulate four conditions based on which the invariant probability measure of the perturbed Markov chain, denoted by $\pi(\varepsilon)$, is derived. These four conditions guarantee the analyticity of $\pi(\varepsilon)$ in ε in a punctured neighborhood of zero. Furthermore, they guarantee that the coefficients of the power series $\pi(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m \pi^{(m)}$ form a geometric sequence and, hence, that there exists a computationally stable updating formula for $\pi(\varepsilon)$, see [2].
In this subsection we only formulate the four conditions and give the main result of the section. The detailed mathematical proofs follow in the next subsection.

Condition 7.1. *The unperturbed Markov chain consists of several (denumerable) ergodic classes and there are no transient states.*

There is an ergodic class for each $\boldsymbol{i} \in \left\{(n_1, n_2),\ n_1, n_2 \in \mathbb{N}\right\}$, i.e., in an ergodic class, the numbers of customers in both queues are fixed. All ergodic classes are identical, and consist of two states, $k \in \{1,2\}$, indicating the queue being served.

Condition 7.2. *The Markov chains corresponding to the ergodic classes of the unperturbed Markov chain are uniformly Lyapunov stable i.e., for each ergodic class there exist a strongly aperiodic state $\alpha \in \{1,2\}$ (with a strictly positive probability on the corresponding diagonal element of the transition matrix $\boldsymbol{I}+\boldsymbol{C}$, with the matrix $\boldsymbol{C}$ given in (7.2)), constants $0 < \delta < 1$ and $b < \infty$, and a Lyapunov function $\boldsymbol{u} = \begin{pmatrix} u_1 & u_2 \end{pmatrix}'$, with $u_i \geq 1$, $i = 1,2$, such that*

$$(\boldsymbol{I}+\boldsymbol{C})\boldsymbol{u} \leq \delta \boldsymbol{u} + b\boldsymbol{e}_\alpha, \tag{7.7}$$

with $\boldsymbol{e}_\alpha$ a vector with 1 in the entry belonging to state α and zero in the other entry.

For the next condition, we first introduce the *aggregated* Markov chain [13, 15, 28], with generator $\boldsymbol{\Gamma}$, given in matrix form as follows

$$\boldsymbol{\Gamma} = \boldsymbol{V}\boldsymbol{G}^{(1)}\boldsymbol{W}, \tag{7.8}$$

with $\boldsymbol{V}$ and $\boldsymbol{W}$ defined as in [2, p. 844]; $\boldsymbol{V}$ (resp., $\boldsymbol{W}$) is a matrix whose rows (columns) correspond to the ergodic classes and its columns (rows) to the states in $\mathscr{S}$. The $\boldsymbol{i}$-th row of $\boldsymbol{V}$ is the invariant measure of the unperturbed Markov chain, given that the process starts in the $\boldsymbol{i}$-th ergodic class, with $\boldsymbol{i} \in \left\{(n_1, n_2),\ n_1, n_2 \in \mathbb{N}\right\}$, i.e.,

$$\boldsymbol{V} = \begin{bmatrix} \tilde{\boldsymbol{C}} & \boldsymbol{0}_{1\times 2} & \boldsymbol{0}_{1\times 2} & \cdots \\ \boldsymbol{0}_{1\times 2} & \tilde{\boldsymbol{C}} & \boldsymbol{0}_{1\times 2} & \cdots \\ \boldsymbol{0}_{1\times 2} & \boldsymbol{0}_{1\times 2} & \tilde{\boldsymbol{C}} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{7.9}$$

with $\tilde{\boldsymbol{C}} = \begin{bmatrix} c_2/(c_1+c_2) & c_1/(c_1+c_2) \end{bmatrix}$. The $\boldsymbol{j}$-th column of $\boldsymbol{W}$ has ones in the components corresponding to the $\boldsymbol{j}$-th ergodic class and zeros in the other components, with $\boldsymbol{j} \in \left\{(n_1, n_2),\ n_1, n_2 \in \mathbb{N}\right\}$, i.e.,

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{1}_{2\times 1} & \boldsymbol{0}_{2\times 1} & \boldsymbol{0}_{2\times 1} & \cdots & \\ \boldsymbol{0}_{2\times 1} & \boldsymbol{1}_{2\times 1} & \boldsymbol{0}_{2\times 1} & \cdots & \\ \boldsymbol{0}_{2\times 1} & \boldsymbol{0}_{2\times 1} & \boldsymbol{1}_{2\times 1} & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \end{bmatrix} \tag{7.10}$$

with $\boldsymbol{1}_{2\times 1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
Hence, for $n_1 \geq 0$, $n_2 \geq 0$, the elements of the generator matrix $\boldsymbol{\Gamma}$ are:

$$\begin{aligned} &\boldsymbol{\Gamma}_{(n_1,n_2),(n_1+1,n_2)} = \lambda_1,\ \boldsymbol{\Gamma}_{(n_1,n_2),(n_1,n_2+1)} = \lambda_2,\ \boldsymbol{\Gamma}_{(n_1+1,n_2),(n_1,n_2)} = \mu_1 \frac{c_2}{c_1+c_2}, \\ &\boldsymbol{\Gamma}_{(n_1,n_2+1),(n_1,n_2)} = \mu_2 \frac{c_1}{c_1+c_2},\ \boldsymbol{\Gamma}_{(n_1,n_2),(n_1,n_2)} = -\left(\lambda_1 + \lambda_2 + \mu_1 \frac{c_2}{c_1+c_2} \mathbb{1}_{\{n_1 \geq 1\}} + \mu_2 \frac{c_1}{c_1+c_2} \mathbb{1}_{\{n_2 \geq 1\}}\right). \end{aligned} \tag{7.11}$$

It is convenient to think of the aggregated Markov chain as the limiting joint queue length process as $\varepsilon \to 0$. In this limit, the server moves infinitely fast between the two queues, making them two independent M/M/1 queues with arrival rates λ_i and service rates $\mu_i \frac{c_1 c_2/c_i}{c_1+c_2}$, $i=1,2$. Based on this remark, one can immediately deduce that the invariant probability measure of the aggregated Markov chain is

$$\bar{\pi}(n_1,n_2) = (1-\tilde{\rho}_1)\tilde{\rho}_1^{n_1}(1-\tilde{\rho}_2)\tilde{\rho}_2^{n_2},\ n_1,n_2 \geq 0, \tag{7.12}$$

with $\tilde{\rho}_i = \frac{\lambda_i c_i(c_1+c_2)}{\mu_i c_1 c_2}$, $i=1,2$.
We are now ready to state the third condition.

Condition 7.3. *The aggregated Markov chain is irreducible and Lyapunov stable, i.e., there exist a strongly aperiodic state $\bar{\alpha} = (n_1,n_2)$ (with a strictly positive probability on the diagonal of the transition matrix $\boldsymbol{I}+\boldsymbol{\Gamma}$, with the matrix $\boldsymbol{\Gamma}$ given in (7.11)), constants $0<\bar{\delta}<1$, $\bar{b}<\infty$ and a Lyapunov function $\bar{\boldsymbol{u}} = \left(\bar{u}_{(n_1,n_2)}\right)_{(n_1,n_2)\in\mathbb{N}^2}$, with elements $\bar{u}_{(n_1,n_2)} \geq 1$, for all $n_1,n_2 \geq 0$, such that*

$$(\boldsymbol{I}+\boldsymbol{\Gamma})\bar{\boldsymbol{u}} \leq \bar{\delta}\bar{\boldsymbol{u}} + \bar{b}\boldsymbol{e}_{\bar{\alpha}}. \tag{7.13}$$

Condition 7.4. *The perturbation matrix $\boldsymbol{G}^{(1)}$ is $\tilde{\boldsymbol{u}}$-bounded (for $\tilde{u}_{\boldsymbol{i}k} = \bar{u}_{\boldsymbol{i}} u_k$, with $\boldsymbol{i} \in \left\{(n_1,n_2),\ n_1,n_2 \in \mathbb{N}\right\}$ and $k=1,2$) or, equivalently,*

$$\parallel \boldsymbol{G}^{(1)} \parallel_{\tilde{\boldsymbol{u}}} := \sup_{\boldsymbol{s}\in\mathscr{S}} \tilde{u}_{\boldsymbol{s}}{}^{-1} \sum_{\bar{\boldsymbol{s}}\in\mathscr{S}} \left|G^{(1)}_{\boldsymbol{s},\bar{\boldsymbol{s}}}\right| \tilde{u}_{\bar{\boldsymbol{s}}} \tag{7.14}$$

is bounded by some constant $g>0$, cf. [2, p. 841].

Note that, because of the repetitive structure of $\boldsymbol{G}^{(0)}$, this assumption implies that $\boldsymbol{P}(\varepsilon)$ is $\tilde{\boldsymbol{u}}$-bounded for all $\varepsilon \geq 0$.

We can now state the main theorem of the section, which is based on [2, p. 845, Theorem 4.1].

Theorem 7.1. *Under Conditions 7.1–7.4, the perturbed Markov chain has a unique invariant probability measure, $\pi(\varepsilon)$, which is an analytic function of ε in a neighborhood of 0,*

$$\pi(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m \pi^{(m)},\ \ \pi^{(m)} = \bar{\pi}\boldsymbol{V}\boldsymbol{U}^m, \tag{7.15}$$

where $\bar{\pi}$ is the invariant probability measure of the aggregated Markov chain, cf. (7.12), and

$$\boldsymbol{U} = \boldsymbol{G}^{(1)}\boldsymbol{H}\left(\boldsymbol{I}+\boldsymbol{G}^{(1)}\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{V}\right), \tag{7.16}$$

$\boldsymbol{V}$ and $\boldsymbol{W}$ are given in (7.9) and (7.10), respectively, and $\boldsymbol{H}$ and $\boldsymbol{\Phi}$ the deviation matrices of the unperturbed and aggregated Markov chains, respectively, are given by

$$\boldsymbol{H} = -\frac{1}{(c_1+c_2)^2}\boldsymbol{G}^{(0)}, \tag{7.17}$$

and

$$\boldsymbol{\Phi} = \sum_{m=0}^{\infty}\left[(\boldsymbol{I}+\boldsymbol{\Gamma})^m - \boldsymbol{\gamma}\right]. \tag{7.18}$$

Here $\boldsymbol{\gamma}$ is the ergodic projection of the aggregated Markov chain, with generator $\boldsymbol{\Gamma}$ given in (7.11), i.e.,

$$\boldsymbol{\gamma} = \lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}(\boldsymbol{I}+\boldsymbol{\Gamma})^m.$$

Remark 7.1. *We do not discuss the radius of convergence of the series in (7.15). Theorem 4.1 of [2] gives a (rather conservative) bound for the analytic region.*

Remark 7.2. *The invariant probability measure of the perturbed Markov chain can be calculated by the updating formula*

$$\pi(\varepsilon) = \pi^{(0)} \left(\boldsymbol{I} - \varepsilon \boldsymbol{U}\right)^{-1}, \tag{7.19}$$

with ε in a neighborhood of 0, cf. [2, p. 845, Theorem 4.1].

Remark 7.3. *In order to calculate the deviation matrix $\boldsymbol{\Phi}$, one may use the following equations*

$$\boldsymbol{\Phi}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Phi} = \boldsymbol{\gamma} - \boldsymbol{I},$$
$$\boldsymbol{\gamma}\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\gamma} = \boldsymbol{0}.$$

We briefly describe two approaches to obtain the deviation matrix $\boldsymbol{\Phi}$: an analytic one involving PGFs and a numerical one. Both approaches require some additional work. The analytic approach, which involves the consideration of generating functions, leads to a boundary value problem for which we can employ Steps 1 and 2 discussed in Section 6. Performing these steps reveals a problem similar to the combinatorial random walk in the quadrant with transitions to the West, North, and South-East, cf. [6, Section 5.2]. In order to obtain the expression for $\boldsymbol{\Phi}$, we need to invert the obtained PGF. A numerical approach is to truncate the state space and solve numerically the corresponding finite system of equations above. We do remark that truncating the state space is a delicate task, since the entries of $\boldsymbol{\Phi}$ corresponding to states far from the origin are unbounded.

7.2 Singular perturbation analysis: verification of the conditions

It remains to prove that Conditions 7.1 - 7.4 are satisfied and also to indicate how the deviation matrix of the unperturbed Markov chain, $\boldsymbol{H}$, is calculated.

Verification of Condition 7.1: As explained in the previous section, this condition follows directly from Equation (7.6).

Verification of Condition 7.2: Obviously, all ergodic classes are identical and contain two states ($k = 1, 2$), thus this condition is trivially met, but for the construction in the remainder it is useful to specify the Lyapunov function used.

First note that the strong aperiodicity follows from the uniformization condition (7.4). We can choose any of the two states as the strongly aperiodic state; in the following we choose $\alpha := 1$. To construct the Lyapunov function first we choose the constants δ and b as $\delta \in \left(1 - c_2, 1 - \frac{c_1 c_2}{c_1 + c_2}\right)$, $b = 1 - \delta + \frac{c_1^2}{c_2}$. Then we can verify that the Lyapunov function

$$\boldsymbol{u} = \begin{bmatrix} 1 \\ 1 + \frac{c_1}{c_2} \end{bmatrix} \tag{7.20}$$

satisfies (7.7). It also follows that, indeed, $\delta \in (0, 1)$, $0 < b < \infty$ and $u_k \geq 1$, $k = 1, 2$.

Verification of Condition 7.3: From the definition of the generator of the aggregated Markov chain, cf. (7.11), and the stability condition (2.1), it is immediately evident that the aggregated Markov chain is ergodic, since it behaves as two independent ergodic M/M/1 queues with arrival rate λ_i and service rate $\frac{\mu_i}{c_i} \frac{c_1 c_2}{c_1 + c_2}$, $i = 1, 2$.

Now by using the uniformization condition (7.4), state $(0, 0)$ is strongly aperiodic i.e., we may choose $\bar{\alpha} = (0, 0)$. We proceed to describe the Lyapunov function $\bar{\boldsymbol{u}}$ and the constants $\bar{\delta} \in (0, 1)$ and $\bar{b}$ which satisfy Condition 7.3. Note that Equation (7.13) is written as follows, for $n_1, n_2 \geq 0$,

$$\begin{aligned}
&\left(1 - \left(\lambda_1 + \lambda_2 + \mu_1 \frac{c_2}{c_1 + c_2} \mathbb{1}_{\{n_1 \geq 1\}} + \mu_2 \frac{c_1}{c_1 + c_2} \mathbb{1}_{\{n_2 \geq 1\}}\right)\right) \bar{u}_{(n_1, n_2)} \\
&+ \lambda_1 \bar{u}_{(n_1+1, n_2)} + \lambda_2 \bar{u}_{(n_1, n_2+1)} + \mu_1 \frac{c_2}{c_1 + c_2} \mathbb{1}_{\{n_1 \geq 1\}} \bar{u}_{(n_1-1, n_2)} + \mu_2 \frac{c_1}{c_1 + c_2} \mathbb{1}_{\{n_2 \geq 1\}} \bar{u}_{(n_1, n_2-1)} \\
&\leq \bar{\delta} \bar{u}_{(n_1, n_2)} + \bar{b} \mathbb{1}_{\{(n_1, n_2) = (0, 0)\}}.
\end{aligned}$$

Solving the above equations with equality, after choosing

$$\bar{u}_{(n_1,n_2)} = \left(\sqrt{\frac{\mu_1 c_2}{\lambda_1(c_1+c_2)}}\right)^{n_1}\left(\sqrt{\frac{\mu_2 c_1}{\lambda_2(c_1+c_2)}}\right)^{n_2}, \tag{7.21}$$

yields the solution for $\bar{\delta}$ and $\bar{b}$. We choose

$$\begin{aligned}\bar{\delta} =& 1-\left(\sqrt{\lambda_1}-\sqrt{\mu_1\frac{c_2}{c_1+c_2}}\right)^2-\left(\sqrt{\lambda_2}-\sqrt{\mu_2\frac{c_1}{c_1+c_2}}\right)^2\\ &+\max\left\{\mu_2\frac{c_1}{c_1+c_2}\left(1-\sqrt{\frac{\lambda_2(c_1+c_2)}{\mu_2 c_1}}\right),\mu_1\frac{c_2}{c_1+c_2}\left(1-\sqrt{\frac{\lambda_1(c_1+c_2)}{\mu_1 c_2}}\right)\right\},\end{aligned}$$

and

$$\bar{b} = 1-\bar{\delta}+\lambda_1\left(\sqrt{\frac{\mu_1 c_2}{\lambda_1(c_1+c_2)}}-1\right)+\lambda_2\left(\sqrt{\frac{\mu_2 c_1}{\lambda_2(c_1+c_2)}}-1\right).$$

Note that due to the uniformization condition (7.4), indeed $\bar{\delta} \in (0,1)$, $0<\bar{b}<\infty$ and $\bar{u}_{(n_1,n_2)} \geq 1$, for all $n_1,n_2 \geq 0$.

Verification of condition 7.4: To verify this assumption, we apply the definition, cf. (7.14), and show that

$$\| \boldsymbol{G}^{(1)} \|_{\tilde{\boldsymbol{u}}} \leq \max\{g_1,g_2\},$$

with $g_1 = \left(\frac{\mu_1 c_2}{\lambda_1(c_1+c_2)}\right)^{-\frac{1}{2}}\left(\mu_1+\frac{\mu_1 c_2+\mu_2 c_1}{c_1+c_2}\right)$ and $g_2 = \left(\frac{\mu_2 c_1}{\lambda_2(c_1+c_2)}\right)^{-\frac{1}{2}}\left(\mu_2+\frac{\mu_1 c_2+\mu_2 c_1}{c_1+c_2}\right)$.
In order to do so, we use the following $\tilde{\boldsymbol{u}}$-norm

$$\tilde{u}_{(n_1,n_2,k)} = \bar{u}_{(n_1,n_2)}u_k,\ (n_1,n_2,k)\in\mathscr{S},$$

with $\bar{u}_{(n_1,n_2)}$ given in (7.21) and u_k given in (7.20).

Derivation of the deviation matrix of the unperturbed Markov chain: It follows from Condition 7.1, that the deviation matrix of the unperturbed Markov chain, $\boldsymbol{H}$, has the following block diagonal structure

$$\boldsymbol{H} = \begin{bmatrix}\boldsymbol{H}_{2\times 2} & \boldsymbol{0}_{2\times 2} & \cdots\\ \boldsymbol{0}_{2\times 2} & \boldsymbol{H}_{2\times 2} & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}, \tag{7.22}$$

with $\boldsymbol{H}_{2\times 2}$ the deviation matrix of each ergodic class of the unperturbed Markov chain, i.e.,

$$\boldsymbol{H}_{2\times 2} = \sum_{j=0}^{\infty}\left[(\boldsymbol{I}+\boldsymbol{C})^j-\boldsymbol{c}\right], \tag{7.23}$$

with $\boldsymbol{C}$ given in (7.2) and $\boldsymbol{c}$ the ergodic projection of the unperturbed Markov chain given as

$$\boldsymbol{c} = \begin{bmatrix}\frac{c_2}{c_1+c_2} & \frac{c_1}{c_1+c_2}\\ \frac{c_2}{c_1+c_2} & \frac{c_1}{c_1+c_2}\end{bmatrix},$$

cf. [31, p. 64, Equation 4.1].
We evaluate (7.23) using the spectral decomposition (eigen-decomposition) of matrices $\boldsymbol{I}+\boldsymbol{C}$ and $\boldsymbol{c}$; the diagonal matrices containing the eigenvalues are $\boldsymbol{D}_{\boldsymbol{I}+\boldsymbol{C}} = \operatorname{diag}\{1,1-(c_1+c_2)\} = \begin{bmatrix}1 & 0\\ 0 & 1-(c_1+c_2)\end{bmatrix}$ and $\boldsymbol{D}_{\boldsymbol{c}} = \operatorname{diag}\{1,0\} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$, respectively, and the corresponding matrix of eigenvectors is $\boldsymbol{M} = \begin{bmatrix}1 & -c_1\\ 1 & c_2\end{bmatrix}$. Naturally, in dimension 2,

both matrices produce the same eigenvectors because $\boldsymbol{c}$ is the ergodic projection of $\boldsymbol{I}+\boldsymbol{C}$. Therefore, Equation (7.23) can be written as

$$\begin{aligned} \boldsymbol{H}_{2\times 2} &= \boldsymbol{M}\left(\sum_{m=0}^{\infty}\left[\boldsymbol{D}_{\boldsymbol{I}+\boldsymbol{C}}^{m}-\boldsymbol{D}_{\boldsymbol{c}}\right]\right)\boldsymbol{M}^{-1} \\ &= -\frac{1}{(c_1+c_2)^2}\boldsymbol{C}. \end{aligned} \tag{7.24}$$

Combining (7.24) and (7.22) yields Equation (7.17).

8 Possible future directions

We have studied a single server two-queue polling model with a random residing time service discipline. More concretely, we considered that customers arrive at the two queues according to two independent Poisson processes. There is a single server that serves both queues with generally distributed service times. The server spends an exponentially distributed amount of time in each queue. After the completion of this residing time, the server instantaneously switches to the other queue, i.e., there is no switch-over time. A service discipline with a random residing time does not satisfy the so-called branching property [29], which complicates significantly the underlying analysis.

For this polling model, we derived the steady-state marginal workload distribution and discussed the complications arising in the calculation of the joint workload distribution. Furthermore, restricting ourselves to the case of exponential service times, we have also calculated the joint queue length distribution using (singular) perturbation analysis. The insights gained for the two-queue polling model, specifically for the derivation of the marginal workload, cf. Section 3, can be also used in the case of N queues, $N > 2$. In addition, one may consider instead of a Poisson input a more general Lévy input. Also, the analysis at hand stands in the case of dependent arrivals streams at the queues.

Another interesting tangent for future research is to develop the framework for the derivation of the bivariate LST of the joint workload distribution, cf. Section 6, or similarly, for the derivation of the bivariate PGF of the joint queue length in the case of exponential service requirements. In particular, the objective in such a setting is to develop an approach for the transformation of Equation (6.4) into a Riemann or Riemann-Hilbert boundary value problem. This requires, that we first choose the zeros of the kernel equation $K(s_1,s_2)$, so as to define a closed smooth contour. Thereafter, we need to show that Equation (6.4) on the contour reduces to the study of an analytic function with a known boundary condition. The main challenge of such an approach lies in the fact that the typical choice of complex conjugate points does not reveal an analytic function, cf. Equation (6.6), thus indicating that we may need to apply a different approach. To this end, an interesting direction would be to extend the framework developed by Fayolle et al. [17], of the systematic use of the *kernel method* using the group of *birational transformations* that leave the kernel equation unchanged. The challenge in our case is that the kernel $K(s_1,s_2)$ does not have the regular structure indicated in [17], but still this does not seem to impose an insuperable obstacle, see, e.g., [23].

Acknowledgments

The authors gratefully acknowledge useful discussions with Offer Kella, Socrates Olympios, and Shelly Zacks. The research of Mayank Saxena was funded by the NWO TOP-C1 project of the Netherlands Organisation for Scientific Research. The research of Onno Boxma was partly done in the framework of the IAP BESTCOM project, funded by the Belgian government; it was also funded by the NWO Gravitation Program NETWORKS of the Dutch government. The work of Stella Kapodistria is supported by the NWO Gravitation Program NETWORKS of the Dutch government.

References

[1] E. Altman. Analysing timed-token ring protocols using the power series algorithm. *In : The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, J. Labetoulle, and J. W. Roberts, editors*, pages 961–971, 1994.

[2] E. Altman, K. E. Avrachenkov, and R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. *Advances in Applied Probability*, 36(3):839–853, 2004.

[3] B. Avi-Itzhak and P. Naor. Some queuing problems with the service station subject to breakdown. *Operations Research*, 11(3):303–320, 1963.

[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27. Cambridge University Press, 1989.

[5] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Applications of polling systems. *Surveys in Operations Research and Management Science (SORMS)*, 16:67–82, 2011.

[6] M. Bousquet-Mélou and M. Mishna. Walks with small steps in the quarter plane. *Contemporary Mathematics*, 520:1–40, 2010.

[7] O. J. Boxma. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, 5(1-3):185–214, 1989.

[8] O. J. Boxma and J. W. Cohen. Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. *Queueing Systems*, 33(1-3):177–204, 1999.

[9] O. J. Boxma and I. A. Kurkova. The $M/G/1$ queue with two service speeds. *Advances in Applied Probability*, 33:520–540, 2001.

[10] J. W. Cohen. Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability*, 10:343–353, 1973.

[11] J. W. Cohen. *The Single Server Queue*. Elsevier, 2012.

[12] J. W. Cohen and O. J. Boxma. *Boundary Value Problems in Queueing System Analysis*, volume 79. Elsevier, 2000.

[13] P. J. Courtois. *Decomposability: Queueing and Computer System Applications*. Academic Press, 2014.

[14] R. de Haan, R. J. Boucherie, and J. C. W. van Ommeren. A polling model with an autonomous server. *Queueing Systems*, 62(3):279–308, 2009.

[15] F. Delebecque. A reduction process for perturbed Markov chains. *SIAM Journal on Applied Mathematics*, 43(2):325–350, 1983.

[16] I. Eliazar and U. Yechiali. Polling under the randomly timed gated regime. *Stochastic Models*, 14(1–2):79–93, 1998.

[17] G. Fayolle, R. Iasnogorodski, and V. Malyshev. *Random Walks in the Quarter-Plane: Algebraic Methods, Boundary Value Problems and Applications*, volume 40 of *Applications of Mathematics*. Springer–Verlag, 1999.

[18] A. Federgruen and L. Green. Queueing systems with service interruptions. *Operations Research*, 34(5):752–768, 1986.

[19] D. Fiems, T. Maertens, and H. Bruneel. Queueing systems with different types of server interruptions. *European Journal of Operational Research*, 188(3):838–845, 2008.

[20] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[21] F. D. Gakhov. *Boundary Value Problems*. Dover Publications, 1990.

[22] D. P. Gaver. A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(1):73–90, 1962.

[23] E. J. Jance van Rensburg, T. Prellberg, and A. Rechnitzer. Partially directed paths in a wedge. *Journal of Combinatorial Theory, Series A*, 115(4):623–650, 2008.

[24] G. Kramer, B. Mukherjee, and G. Pesavento. IPACT a dynamic protocol for an Ethernet PON (EPON). *IEEE Communications Magazine*, 40(2):74–80, 2002.

[25] A. Krishnamoorthy, P. Pramod, and S. Chakravarthy. Queues with interruptions: a survey. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research*, 22(1):290–320, 2014.

[26] H. Levy and M. Sidi. Polling models: applications, modelling and optimization. *IEEE Transactions on Communications*, 38(10):1750–1760, 1990.

[27] L. W. Miller. *Alternating Priorities in Multi-Class Queues*. PhD thesis, Cornell University, Ithaca, N.Y., 1964.

[28] A. A. Pervozvanskii and V. G. Gaitsgori. *Theory of Suboptimal Decisions: Decomposition and Aggregation*, volume 12. Springer Science & Business Media, 2013.

[29] J. A. C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(4):409–426, 1993.

[30] K. Sigman. Appendix: a primer on heavy-tailed distributions. *Queueing Systems*, 33(1-3):261–275, 1999.

[31] F. M. Spieksma. *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*. PhD thesis, Leiden University, 1990.

[32] H. Takagi. Queuing analysis of polling models. *ACM Computing Surveys (CSUR)*, 20(1):5–28, 1988.

[33] H. Takagi. Application of polling models to computer networks. *Computer Networks and ISDN systems*, 22(3):193–211, 1991.

[34] H. Takagi. *Queueing Analysis, Vol. 1: Vacation and Priority Systems*. North-Holland, 1991.

[35] H. Takagi. Analysis and application of polling models. In *Performance Evaluation: Origins and Directions. In G. Haring, C. Lindemann, and M. Reiser, editors*, pages 423–442. Springer, 2000.

[36] T. Takine and B. Sengupta. A single server queue with service interruptions. *Queueing Systems*, 26(3):285–300, 1997.

[37] K. Thiruvengadam. Queuing with breakdowns. *Operations Research*, 11(1):62–71, 1963.

[38] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.

[39] H. White and L. S. Christie. Queuing with preemptive priorities or with breakdown. *Operations Research*, 6(1):79–95, 1958.

[40] J. Xie, M. J. Fischer, and C. M. Harris. Workload and waiting time in a fixed-time loop system. *Computers & Operations Research*, 24(8):789–803, 1997.

[41] A. P. Zwart. *Queueing Systems with Heavy Tails*. PhD thesis, Eindhoven University of Technology, 2001.