

# HEAVY-TRAFFIC ANALYSIS OF $K$ -LIMITED POLLING SYSTEMS

M.A.A. BOON

*Eurandom and Department of Mathematics and Computer Science,  
Eindhoven University of Technology,  
P.O. Box 513,  
5600MB Eindhoven,  
The Netherlands  
E-mail: [marko@win.tue.nl](mailto:marko@win.tue.nl)*

E.M.M. WINANDS

*University of Amsterdam,  
Korteweg-de Vries Institute for Mathematics,  
Science Park 904, 1098 XH Amsterdam,  
The Netherlands  
E-mail: [e.m.m.winands@uva.nl](mailto:e.m.m.winands@uva.nl)*

In this paper, we study a two-queue polling model with zero switchover times and  $k$ -limited service (serve at most  $k_i$  customers during one visit period to queue  $i$ ,  $i = 1, 2$ ) in each queue. The arrival processes at the two queues are Poisson, and the service times are exponentially distributed. By increasing the arrival intensities until one of the queues becomes critically loaded, we derive exact heavy-traffic limits for the joint queue-length distribution using a singular-perturbation technique. It turns out that the number of customers in the stable queue has the same distribution as the number of customers in a vacation system with Erlang- $k_2$  distributed vacations. The queue-length distribution of the critically loaded queue, after applying an appropriate scaling, is exponentially distributed. Finally, we show that the two queue-length processes are independent in heavy traffic.

## 1. INTRODUCTION

This paper considers a two-queue  $k$ -limited polling model with exponentially distributed service times and zero switchover times. Under the  $k$ -limited strategy the server continues working until either a predefined number of  $k_i$  customers is served at queue  $i$  or until the queue becomes empty, whichever occurs first. The interest for this model is fueled by a number of different application areas. That is, the  $k$ -limited strategy did not only prove its merit in communication systems (see, e.g., [5,8]), but also in the field of logistics (see, e.g. [28]), and vehicle-actuated traffic signals [4]. In the present paper, we consider the heavy-traffic scenario, in which one of the queues becomes critically loaded with the other queue remaining stable.

Although the number of papers on polling systems is impressive, hardly any exact results for polling systems with the  $k$ -limited service policy have been obtained. This can be explained by the fact that the  $k$ -limited strategy does not satisfy a well-known branching property for polling systems, independently discovered by Fuhrmann [15] and Resing [26], which significantly increases the analytical complexity. For this reason, most of the papers in the polling literature focus on branching-type service disciplines, like exhaustive service (serve all customers in a queue until it is empty, before switching to the next queue) or gated service (serve all customers present at the server's arrival to a queue, before switching to the next queue). Groenendijk [16] and Ibe [17] give an explicit Laplace–Stieltjes Transform for the waiting-time distribution in a two-queue 1-limited/exhaustive system. For two-queue systems where both queues are served according to the 1-limited discipline, the problem of finding the queue-length distribution can be shown to translate into a boundary value problem [6,7,11,13]. For general  $k$ , an exact evaluation for the queue-length distribution is known in two-queue exhaustive/ $k$ -limited systems (see [21,24,25,28]). The situation where both queues follow the  $k$ -limited discipline has not been solved yet.

When studying the literature on heavy-traffic results for polling systems, a remarkable observation is that distributional heavy-traffic limits have only been rigorously proven for systems with branching-type service disciplines (there is only exception as described later). Even for these systems, the proof is limited to exhaustive systems consisting of two queues [9,10] or systems with Poisson arrivals [27]; all other results are based on conjectures. That is, a proof for the general setting is already a challenge for almost 20 years since the publication of the Coffman/Puhalskii/Reiman papers [9,10]. Moreover, the only paper analyzing the heavy-traffic behavior for the  $k$ -limited discipline is the paper of Lee [21] for the two-queue exhaustive/ $k$ -limited system without setup times. More specifically, he studies the limiting regime where the exhaustive queue remains stable, whereas the  $k$ -limited queue becomes critically loaded in the limit. Lee [21] uses results for this system under stable conditions as a basis to derive heavy-traffic asymptotics by exploiting the fact that significant simplifications result as the load increases. However, corresponding steady-state results are not available for the  $k$ -limited system under consideration in the current paper, which impels us to look for a different family of techniques.

The main contribution of the present paper is that we derive heavy-traffic asymptotics for  $k$ -limited polling models via the singular-perturbation technique. That is, by increasing the arrival intensities until one of the queues becomes critically loaded, we derive heavy-traffic limits for the joint queue-length distribution in a two-queue  $k$ -limited polling model. In this way, we derive the lowest-order asymptotic to the joint queue-length distribution in terms of a small positive parameter measuring the closeness of the system to instability.

Furthermore, the results obtained in the present paper provide new insights into the heavy-traffic behavior of  $k$ -limited polling systems. It is shown that the number of customers in the stable queue has the same distribution as the number of customers in a vacation system with Erlang- $k_2$  distributed vacations, while the scaled queue-length distribution of the critically loaded queue is exponentially distributed. Finally, we prove that the two queue-length processes are independent in heavy traffic. The singular-perturbation technique can also be extended to an  $N$ -queue system ( $N \geq 2$ ) with one queue becoming critically loaded. In this limiting regime, the stable queues have the same joint distribution of a  $k$ -limited polling model with  $N - 1$  queues and an extended switchover time. These results do not only generalize those derived in [21] (where the special two-queue case is studied in which only one queue is served according to the  $k$ -limited service policy), but are also obtained via a fundamentally different singular-perturbation approach. We would like to note that the observed heavy-traffic behavior is different from the behavior seen for branching policies

in which the cycle time normally converges to a Gamma distribution and the individual workloads follow from an averaging principle.

Finally, to our opinion, the novel application of the singular perturbation technique to polling systems is interesting in itself. That is, the merits of this technique are in its intrinsic simplicity and its intuitively appealing derivation, although it requires some distributional assumptions. The singular perturbation framework is known as a powerful tool to determine asymptotics in all kinds of queueing models, but it has never been applied to polling systems. We would like to refer to Knessl and Tier [19] for an excellent survey of applications of the perturbation technique to queueing models. It is noteworthy that our paper is inspired by the manner in which Morrison and Borst [23] apply this technique to a model with interacting queues.

The paper is structured as follows. In the next section, we introduce the model and notation. In Section 3, we apply a perturbation technique to study the system under heavy-traffic conditions and derive the limiting scaled joint queue-length distribution. In Section 4, we interpret the results and give some suggestions on further research. The appendices contain some lengthy derivations required for the analysis in Section 3.

## 2. MODEL DESCRIPTION AND NOTATION

We consider a polling model consisting of two queues,  $Q_1$  and  $Q_2$ , that are alternately visited by a single server. Throughout this paper, the subscript  $i$  will always be used to refer to one of the queues, meaning that it always takes on the values 1 or 2. When a server arrives at  $Q_i$ , it serves at most  $k_i$  customers. When  $k_i$  customers have been served or  $Q_i$  becomes empty, whichever occurs first, the server switches to the other queue. We assume that switching from one queue to the other requires no time. If the other queue turns out to be empty, the server switches back and serves, again, at most  $k_i$  customers. If both queues are empty, the server waits until the first arrival and switches to the corresponding queue (say,  $Q_j$ ) to start another visit period of at most  $k_j$  customers,  $j = 1, 2$ . Customers arrive at  $Q_i$  according to a Poisson process with intensity  $\lambda_i$ . We assume that the service times of customers in  $Q_i$  are independent and exponentially distributed with parameter  $\mu_i$ . We denote the load of the system by  $\rho = \rho_1 + \rho_2$ , where  $\rho_i = \lambda_i/\mu_i$ . For a polling model without switchover times, the stability condition is  $\rho < 1$  [12,14]. Furthermore, we assume that

$$\frac{\lambda_1}{k_1} < \frac{\lambda_2}{k_2}. \quad (2.1)$$

This assumption is discussed in more detail in the next section.

The number of customers in  $Q_i$  at time  $t$ ,  $t \geq 0$ , is denoted by  $N_i(t)$ . In order to describe the queue-length process as a Markov process, we use the approach of Blanc [3], introducing a supplementary variable  $H(t)$  which takes on values  $1, 2, \dots, k_1 + k_2$ . The variable  $H(t)$  is used to determine the server position ( $Q_1$  or  $Q_2$ ) at time  $t$  and the number of customers that can be served before the server has to switch to the next queue. When  $1 \leq H(t) \leq k_1$ , this means that the server is serving  $Q_1$  at time  $t$ , and that the customer in service is the  $H(t)$ -th customer being served during the present visit period. If  $k_1 + 1 \leq H(t) \leq k_1 + k_2$ , the server is serving the  $(H(t) - k_1)$ -th customer in  $Q_2$ . Now,  $(N_1(t), N_2(t), H(t))$  is a Markov process. Assuming that  $\rho < 1$ , define the stationary probabilities  $p(n_1, n_2, h) := \lim_{t \rightarrow \infty} \mathbb{P}(N_1(t) = n_1, N_2(t) = n_2, H(t) = h)$ , and define the steady-state queue lengths  $N_i$ .

### 3. ANALYSIS

We study the heavy-traffic limit of the joint queue-length process  $(N_1, N_2)$  by increasing the arrival rate  $\lambda_2$ , while keeping  $\lambda_1$  fixed. When  $\rho$  tends to 1, Assumption (2.1) implies that  $Q_2$  will become critically loaded, whereas  $Q_1$  remains stable due to the fact that at most  $k_2$  customers are served during each visit period at  $Q_2$ . In case  $\lambda_1/k_1 = \lambda_2/k_2$  in the limit, both queues would become critically loaded simultaneously and the system behavior is different from the limiting behavior found in the present paper. In fact, the limiting queue-length behavior for that specific case remains an open problem. We discuss this topic briefly in Section 4.

We use a single-perturbation method to find the queue-length distributions in heavy-traffic ( $\rho \uparrow 1$ ). First, we write down the balance equations of our model and apply a perturbation to the arrival rate of  $Q_2$  to these equations, in the case that this queue is close to becoming critically loaded. By solving the system of balance equations for the lowest (zeroth)-order terms, we find the queue-length distribution of the stable queue,  $Q_1$ . By analyzing the equations for the first-order and second-order terms, we also obtain a differential equation for the scaled number of customers in  $Q_2$ , which we can solve to show that this number converges to an exponential distribution.

#### 3.1. Balance Equations

The balance equations for a polling model with exponentially distributed service times and  $k$ -limited service at each of the queues are given by Blanc [3]. For completeness, we present these equations for our two-queue model below.

$$\begin{aligned}
 (\lambda_1 + \lambda_2 + \mu_2)p(0, n_2, k_1 + 1) &= \lambda_2 p(0, n_2 - 1, k_1 + 1) + \mu_2 p(0, n_2 + 1, k_1 + k_2) \\
 &+ \sum_{h=1}^{k_1} \mu_1 p(1, n_2, h), \tag{3.1a}
 \end{aligned}$$

$$(\lambda_1 + \lambda_2 + \mu_2)p(0, n_2, h_2) = \lambda_2 p(0, n_2 - 1, h_2) + \mu_2 p(0, n_2 + 1, h_2 - 1), \tag{3.1b}$$

$$(\lambda_1 + \lambda_2 + \mu_1)p(1, n_2, 1) = \lambda_2 p(1, n_2 - 1, 1) + \mu_2 p(1, n_2 + 1, k_1 + k_2), \tag{3.1c}$$

$$(\lambda_1 + \lambda_2 + \mu_1)p(1, n_2, h_1) = \lambda_2 p(1, n_2 - 1, h_1) + \mu_1 p(2, n_2, h_1 - 1), \tag{3.1d}$$

$$\begin{aligned}
 (\lambda_1 + \lambda_2 + \mu_1)p(n_1 + 1, n_2, 1) &= \lambda_1 p(n_1, n_2, 1) + \lambda_2 p(n_1 + 1, n_2 - 1, 1) \\
 &+ \mu_2 p(n_1 + 1, n_2 + 1, k_1 + k_2), \tag{3.1e}
 \end{aligned}$$

$$\begin{aligned}
 (\lambda_1 + \lambda_2 + \mu_1)p(n_1 + 1, n_2, h_1) &= \lambda_1 p(n_1, n_2, h_1) + \lambda_2 p(n_1 + 1, n_2 - 1, h_1) \\
 &+ \mu_1 p(n_1 + 2, n_2, h_1 - 1), \tag{3.1f}
 \end{aligned}$$

$$\begin{aligned}
 (\lambda_1 + \lambda_2 + \mu_2)p(n_1, n_2, k_1 + 1) &= \lambda_1 p(n_1 - 1, n_2, k_1 + 1) + \lambda_2 p(n_1, n_2 - 1, k_1 + 1) \\
 &+ \mu_1 p(n_1 + 1, n_2, k_1), \tag{3.1g}
 \end{aligned}$$

$$\begin{aligned}
 (\lambda_1 + \lambda_2 + \mu_2)p(n_1, n_2, h_2) &= \lambda_1 p(n_1 - 1, n_2, h_2) + \lambda_2 p(n_1, n_2 - 1, h_2) \\
 &+ \mu_2 p(n_1, n_2 + 1, h_2 - 1), \tag{3.1h}
 \end{aligned}$$

for  $n_1 = 1, 2, \dots; n_2 = 2, 3, \dots; h_1 = 2, 3, \dots, k_1$ , and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . Note that (3.1a)–(3.1h) are *not all* balance equations. We have omitted all equations for  $n_2 = 0$  and  $n_2 = 1$ , since it will turn out that these do not play a role after the perturbation. The intuitive explanation is that  $N_2(t)$  will tend to infinity as  $Q_2$  becomes critically loaded and the probabilities  $p(n_1, n_2, h)$  become negligible for low values of  $n_2$ .

### 3.2. Perturbation

From the stability condition we have that the system becomes unstable as  $\lambda_2/\mu_2 \uparrow 1 - \lambda_1/\mu_1$ , which means that the arrival rate  $\lambda_2$  approaches  $\mu_2(1 - \lambda_1/\mu_1)$ . Therefore we will assume that

$$\lambda_2 = \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) - \delta\omega, \quad \omega > 0, 0 < \delta \ll 1. \tag{3.2}$$

At the end of this section we will take an appropriate choice for the constant  $\omega$ , which will influence the limit of the scaled queue length in  $Q_2$ .

Let  $\xi = \delta n_2$ , and

$$p(n_1, \xi/\delta, h) = \delta\phi_{n_1, h}(\xi, \delta), \quad 0 < \xi = \mathcal{O}(1), \quad h = 1, 2, \dots, k_1 + k_2. \tag{3.3}$$

Note that once  $\omega$  is chosen,  $\delta$  and the scaled variable  $\xi$  will be uniquely defined. The next step is to substitute (3.2) and (3.3) in the balance Eqs. (3.1a)–(3.1h), and take the Taylor series expansion with respect to  $\delta$ . For reasons of compactness, we only show the intermediate results for Eq. (3.1h) as an illustration:

$$\begin{aligned} & (\lambda_1 + \mu_2)\phi_{n_1, h_2}(\xi, \delta) - \lambda_1\phi_{n_1-1, h_2}(\xi, \delta) - \mu_2\phi_{n_1, h_2-1}(\xi + \delta, \delta) \\ & \quad = \left( \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) - \delta\omega \right) (\phi_{n_1, h_2}(\xi - \delta, \delta) - \phi_{n_1, h_2}(\xi, \delta)). \end{aligned}$$

Taking the Taylor series yields:

$$\begin{aligned} & (\lambda_1 + \mu_2)\phi_{n_1, h_2}(\xi, \delta) - \lambda_1\phi_{n_1-1, h_2}(\xi, \delta) \\ & \quad - \mu_2 \left( \phi_{n_1, h_2-1}(\xi, \delta) + \delta \frac{\partial \phi_{n_1, h_2-1}(\xi, \delta)}{\partial \xi} + \frac{\delta^2}{2} \frac{\partial^2 \phi_{n_1, h_2-1}(\xi, \delta)}{\partial \xi^2} \right) \\ & \quad = - \left( \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) - \delta\omega \right) \left( \delta \frac{\partial \phi_{n_1, h_2}(\xi, \delta)}{\partial \xi} - \frac{\delta^2}{2} \frac{\partial^2 \phi_{n_1, h_2}(\xi, \delta)}{\partial \xi^2} \right) + \mathcal{O}(\delta^3). \tag{3.4} \end{aligned}$$

Note that  $\lambda_2$  (or  $\mu_2(1 - \frac{\lambda_1}{\mu_1}) - \delta\omega$  after the substitution) only plays a role in this equation for  $\mathcal{O}(\delta)$  terms and higher. It is readily verified that this is the case for all balance equations. We now expand in powers of  $\delta$ , and let

$$\phi_{n_1, h}(\xi, \delta) = \phi_{n_1, h}^{(0)}(\xi) + \delta\phi_{n_1, h}^{(1)}(\xi) + \mathcal{O}(\delta^2). \tag{3.5}$$

One would expect that also  $\delta^2\phi_{n_1, h}^{(2)}(\xi)$  should be introduced in order to analyze the second-order terms. However, Proposition 3.3 implies that the above power expansion suffices. We also define the corresponding generating functions

$$\tilde{Q}_h(z, \xi, \delta) := \sum_{n_1=0}^{\infty} \phi_{n_1, h}(\xi, \delta) z^{n_1}, \quad \tilde{Q}_h^{(j)}(z, \xi) := \sum_{n_1=0}^{\infty} \phi_{n_1, h}^{(j)}(\xi) z^{n_1}, \quad j = 0, 1, 2, \dots$$

In the next subsections, we first equate the lowest-order terms of the resulting equations to find an expression for (the generating function of)  $\phi_{n, h}^{(0)}(\xi)$ , and subsequently we equate the first-order and second-order terms to find the scaled queue-length distribution of  $Q_2$ .

**3.3. Equating the Lowest-Order Terms**

Equating the lowest-order terms of the balance equations, after substituting (3.2), (3.3), and (3.5), results in the following equations:

$$(\lambda_1 + \mu_2)\phi_{0,k_1+1}^{(0)}(\xi) = \mu_2\phi_{0,k_1+k_2}^{(0)}(\xi) + \sum_{h=1}^{k_1} \mu_1\phi_{1,h}^{(0)}(\xi), \tag{3.6a}$$

$$(\lambda_1 + \mu_2)\phi_{0,h_2}^{(0)}(\xi) = \mu_2\phi_{0,h_2-1}^{(0)}(\xi), \tag{3.6b}$$

$$(\lambda_1 + \mu_1)\phi_{1,1}^{(0)}(\xi) = \mu_2\phi_{1,k_1+k_2}^{(0)}(\xi), \tag{3.6c}$$

$$(\lambda_1 + \mu_1)\phi_{1,h_1}^{(0)}(\xi) = \mu_1\phi_{2,h_1-1}^{(0)}(\xi), \tag{3.6d}$$

$$(\lambda_1 + \mu_1)\phi_{n_1+1,1}^{(0)}(\xi) = \lambda_1\phi_{n_1,1}^{(0)}(\xi) + \mu_2\phi_{n_1+1,k_1+k_2}^{(0)}(\xi), \tag{3.6e}$$

$$(\lambda_1 + \mu_1)\phi_{n_1+1,h_1}^{(0)}(\xi) = \lambda_1\phi_{n_1,h_1}^{(0)}(\xi) + \mu_1\phi_{n_1+2,h_1-1}^{(0)}(\xi), \tag{3.6f}$$

$$(\lambda_1 + \mu_2)\phi_{n_1,k_1+1}^{(0)}(\xi) = \lambda_1\phi_{n_1-1,k_1+1}^{(0)}(\xi) + \mu_1\phi_{n_1+1,k_1}^{(0)}(\xi), \tag{3.6g}$$

$$(\lambda_1 + \mu_2)\phi_{n_1,h_2}^{(0)}(\xi) = \lambda_1\phi_{n_1-1,h_2}^{(0)}(\xi) + \mu_2\phi_{n_1,h_2-1}^{(0)}(\xi), \tag{3.6h}$$

for  $n_1 = 1, 2, \dots; h_1 = 2, 3, \dots, k_1$ , and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ .

Note that  $\sum_{n_1=0}^{\infty} \sum_{h=1}^{k_1+k_2} \phi_{n_1,h}^{(0)}(\xi) \neq 1$ . For this reason we introduce  $P_0(\xi)$  and  $\pi_{n_1,h}^{(0)}$ , with

$$\phi_{n_1,h}^{(0)}(\xi) = \pi_{n_1,h}^{(0)}P_0(\xi), \quad \text{and} \quad \sum_{n_1=0}^{\infty} \sum_{h=1}^{k_1+k_2} \pi_{n_1,h}^{(0)} = 1, \tag{3.7}$$

for  $n_1 = 0, 1, 2, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$ . Careful inspection of these balance equations reveals that Eqs. (3.6a)–(3.6h) describe the behavior of a single-server vacation queue with the following properties:

- P1. The arrival process is Poisson with intensity  $\lambda_1$ ,
- P2. The service times are exponentially distributed with mean  $1/\mu_1$ ,
- P3. The service discipline is  $k$ -limited service with service limit  $k_1$ ,
- P4. The vacations are Erlang- $k_2$  distributed with parameter  $\mu_2$ ,
- P5. Whenever the server finds the system empty upon return from a vacation, it immediately starts another vacation.

This system has been studied in the literature (cf. [22]) and, in general, no closed-form expressions for the steady-state queue-length probabilities can be obtained. However, it is possible to find the probability generating function (PGF) of the queue-length distribution. Define

$$\tilde{L}^{(0)}(z) := \sum_{h=1}^{k_1+k_2} \tilde{L}_h^{(0)}(z), \quad \text{where} \quad \tilde{L}_h^{(0)}(z) := \sum_{n=0}^{\infty} \pi_{n,h}^{(0)}z^n, \quad h = 1, \dots, k_1 + k_2,$$

and

$$\tilde{G}(z) := \frac{\mu_1}{\lambda_1(1-z) + \mu_1}, \quad \tilde{H}(z) := \frac{\mu_2}{\lambda_1(1-z) + \mu_2}. \tag{3.8}$$

It is easily seen that  $\tilde{G}(z)$  and  $\tilde{H}(z)$  are the PGFs of the number of arrivals during respectively one service, and during one stage of the vacation (which consists of  $k_2$  exponential

stages). It follows from (3.6a)–(3.6h) that

$$\tilde{L}_{k_1+k_2}^{(0)}(z) = \frac{\tilde{H}(z)^{k_2} \left[ \pi_{0,k_1+k_2}^{(0)} \left( 1 - \left( \tilde{G}(z)/z \right)^{k_1} \right) + \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} \pi_{1,h}^{(0)} \left( 1 - \left( \tilde{G}(z)/z \right)^{k_1-h} \right) \right]}{1 - \left( \tilde{G}(z)/z \right)^{k_1} \tilde{H}(z)^{k_2}}, \tag{3.9}$$

$$\tilde{L}_{h_1}^{(0)}(z) = \frac{z\mu_2}{\mu_1} \left( \tilde{G}(z)/z \right)^{h_1} \left[ \tilde{L}_{k_1+k_2}^{(0)}(z) - \pi_{0,k_1+k_2}^{(0)} \right] - \sum_{h=1}^{h_1-1} \pi_{1,h}^{(0)} z \left( \tilde{G}(z)/z \right)^{h_1-h}, \tag{3.10}$$

$$\tilde{L}_{h_2}^{(0)}(z) = \tilde{H}(z)^{h_2-k_1} \left[ \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} \pi_{1,h}^{(0)} + \pi_{0,k_1+k_2}^{(0)} + \frac{\mu_1}{z\mu_2} \tilde{L}_{k_1}^{(0)}(z) \right], \tag{3.11}$$

for  $h_1 = 1, \dots, k_1$  and  $h_2 = k_1 + 1, \dots, k_1 + k_2 - 1$ . A derivation of (3.9)–(3.11) can be found in Appendix A. These equations still contain  $k_1$  unknowns:  $\pi_{1,1}^{(0)}, \pi_{1,2}^{(0)}, \dots, \pi_{1,k_1-1}^{(0)}$ , and  $\pi_{0,k_1+k_2}^{(0)}$ . See Appendix A for more details on how to eliminate them using Rouché’s Theorem. Foregoing the derivation of the limiting behavior of  $Q_2$  we already would like to mention that these unknowns do not play a role therein.

We conclude from equating the lowest-order terms of the balance Eqs. (3.1a)–(3.1h), after substituting (3.2), (3.3), and (3.5), that

$$\sum_{n_1=0}^{\infty} \sum_{h=1}^{k_1+k_2} \phi_{n_1,h}^{(0)}(\xi) z^{n_1} = \tilde{L}^{(0)}(z) P_0(\xi), \tag{3.12}$$

where  $P_0(\xi)$  still has to be determined. Consequently, in heavy traffic, the queue length of the stable queue ( $Q_1$ ) has the same distribution as the queue length in a vacation system with Erlang( $k_2$ ) distributed vacations with parameter  $\mu_2$ , exponential service times with parameter  $\mu_1$ , and  $k_1$ -limited service. The assumption that we have Poisson arrivals and first-come-first-served service implies that we can use the distributional form of Little’s law to obtain (the Laplace–Stieltjes transform of) the waiting-time distribution of customers in  $Q_1$  (see, e.g., Keilson and Servi [18]).

**Remark 3.1.** In this paper, we assume that  $\lambda_1/k_1 < \lambda_2/k_2$ , causing  $Q_2$  to become critically loaded when  $\lambda_2$  is being increased. We have implicitly used this assumption when solving the balance Eqs. (3.1a)–(3.1h). It is well known that the vacation system described by these equations is stable if and only if

$$\lambda_1 \mathbb{E}[C] < k_1, \tag{3.13}$$

where  $\mathbb{E}[C]$  is the mean cycle time, that is, the mean length of one visit period plus one vacation. Denoting the length of a vacation by  $S$ , we have  $\mathbb{E}[C] = \mathbb{E}[S]/(1 - \rho_1) = k_2/(\mu_2(1 - \lambda_1/\mu_1))$ . When substituting this in (3.13), we indeed obtain exactly the same inequality as (2.1) after substituting (3.2) and letting  $\delta \downarrow 0$ .

**Remark 3.2.** Another interesting observation is that one could consider more general ways of varying the arrival rates in order to let  $Q_2$  become critically loaded. To this end,

we introduce  $\lambda_1^*$  and  $\lambda_2^*$  such that  $\lambda_1^*/\mu_1 + \lambda_2^*/\mu_2 = 1$ . Additionally, we assume that

$$\frac{\lambda_1^*}{k_1} < \frac{1}{\frac{k_1}{\mu_1} + \frac{k_2}{\mu_2}}, \text{ or equivalently: } \frac{\lambda_2^*}{k_2} > \frac{1}{\frac{k_1}{\mu_1} + \frac{k_2}{\mu_2}}. \tag{3.14}$$

We now let  $\lambda_1 \rightarrow \lambda_1^*$  and  $\lambda_2 \rightarrow \lambda_2^*$  for  $\delta \downarrow 0$ , with

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} = 1 - \delta\omega^*, \quad \omega^* > 0, 0 < \delta \ll 1. \tag{3.15}$$

Any arbitrary way in which we let  $\lambda_1$  and  $\lambda_2$  approach respectively  $\lambda_1^*$  and  $\lambda_2^*$ , for  $\delta \downarrow 0$ , will cause  $Q_2$  to become critically loaded (because of assumption (3.14)). All results obtained in this paper will still be valid, by choosing  $\omega^* = \omega/\mu_2$ .

### 3.4. Equating the First-Order Terms

In this section, we study and solve the system of equations that results from equating the first-order terms of the perturbed balance equations. For notational reasons, we define

$$\psi_{n_1,h}^{(1)}(\xi) := \phi_{n_1,h}^{(1)}(\xi) + \phi_{n_1,h}^{\prime(0)}(\xi), \quad \text{where} \quad \phi_{n_1,h}^{\prime(0)}(\xi) := \frac{d\phi_{n_1,h}^{(0)}(\xi)}{d\xi},$$

for  $n_1 = 0, 1, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$ . The resulting set of equations for the probabilities  $\phi_{n_1,h}^{(1)}(\xi)$  is given below.

$$(\lambda_1 + \mu_2)\phi_{0,k_1+1}^{(1)}(\xi) = \mu_2\psi_{0,k_1+k_2}^{(1)}(\xi) + \sum_{h=1}^{k_1} \mu_1\phi_{1,h}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{0,k_1+1}^{\prime(0)}(\xi), \tag{3.16a}$$

$$(\lambda_1 + \mu_2)\phi_{0,h_2}^{(1)}(\xi) = \mu_2\psi_{0,h_2-1}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{0,h_2}^{\prime(0)}(\xi), \tag{3.16b}$$

$$(\lambda_1 + \mu_1)\phi_{1,1}^{(1)}(\xi) = \mu_2\psi_{1,k_1+k_2}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{1,1}^{\prime(0)}(\xi), \tag{3.16c}$$

$$(\lambda_1 + \mu_1)\phi_{1,h_1}^{(1)}(\xi) = \mu_1\phi_{2,h_1-1}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{1,h_1}^{\prime(0)}(\xi), \tag{3.16d}$$

$$(\lambda_1 + \mu_1)\phi_{n_1+1,1}^{(1)}(\xi) = \lambda_1\phi_{n_1,1}^{(1)}(\xi) + \mu_2\psi_{n_1+1,k_1+k_2}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{n_1+1,1}^{\prime(0)}(\xi), \tag{3.16e}$$

$$(\lambda_1 + \mu_1)\phi_{n_1+1,h_1}^{(1)}(\xi) = \lambda_1\phi_{n_1,h_1}^{(1)}(\xi) + \mu_1\phi_{n_1+2,h_1-1}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{n_1+1,h_1}^{\prime(0)}(\xi), \tag{3.16f}$$

$$(\lambda_1 + \mu_2)\phi_{n_1,k_1+1}^{(1)}(\xi) = \lambda_1\phi_{n_1-1,k_1+1}^{(1)}(\xi) + \mu_1\phi_{n_1+1,k_1}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{n_1,k_1+1}^{\prime(0)}(\xi), \tag{3.16g}$$

$$(\lambda_1 + \mu_2)\phi_{n_1,h_2}^{(1)}(\xi) = \lambda_1\phi_{n_1-1,h_2}^{(1)}(\xi) + \mu_2\psi_{n_1,h_2-1}^{(1)}(\xi) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi_{n_1,h_2}^{\prime(0)}(\xi), \tag{3.16h}$$

for  $n_1 = 1, 2, \dots; h_1 = 2, 3, \dots, k_1$ , and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . The solution to this system of equations, in terms of generating functions, can be found in Appendix A. This solution



is used to derive the following relation:

$$\frac{\lambda_1}{\mu_1} \sum_{n=0}^{\infty} \sum_{h=1}^{k_1+k_2} \phi_{n,h}^{(1)}(\xi) - \sum_{n=0}^{\infty} \sum_{h_1=1}^{k_1} \phi_{n,h_1}^{(1)}(\xi) = \frac{\lambda_1 \mu_2}{\mu_1^2} P_0'(\xi), \tag{3.17}$$

which turns out to play a key role in determining the heavy-traffic limit of the joint queue-length distribution (see the next section).

### 3.5. Equating the Second-Order Terms

In order to find an expression for  $P_0(\xi)$  and, consequently, solve (3.12), we consider the Taylor series of all perturbed balance equations. In this section we show that, fortunately, we only need to consider the *sum* of all these equations (such as Eq. (3.4)) over all  $n_1 = 0, 1, 2, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$ , and we consecutively consider the  $\mathcal{O}(1)$ ,  $\mathcal{O}(\delta)$ , and  $\mathcal{O}(\delta^2)$  terms. Using the results we have obtained so far, we prove that:

1. all  $\mathcal{O}(1)$ , that is, all  $\phi_{n_1,h}(\xi, \delta)$ , terms cancel immediately;
2. the  $\mathcal{O}(\delta)$  terms cancel after expanding  $\phi_{n_1,h}(\xi, \delta)$  in powers of  $\delta$  (i.e., substituting (3.5));
3. the equation that results from equating the  $\mathcal{O}(\delta^2)$  terms, can be solved to find  $P_0(\xi)$ .

The above three results are proven in Propositions 3.3, 3.4 and 3.5.

**Proposition 3.3.** *After taking the summation over all  $n_1 = 0, 1, 2, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$  of the Taylor series of all perturbed balance equations, the  $\mathcal{O}(1)$  terms cancel.*

PROOF: We can follow the generating function approach, used in Section 3.3, but replace the probabilities  $\pi_{n_1,h}^{(0)}$ , by  $\phi_{n_1,h}(\xi, \delta)$ . This results in the same set of equations as (3.9)–(3.11), but with terms  $\tilde{Q}_h(z, \xi, \delta)$  and  $\phi_{n_1,h}(\xi, \delta)$  instead of  $\tilde{L}_h^{(0)}(z)$  and  $\pi_{n_1,h}^{(0)}$ . Substituting  $z = 1$  yields:

$$\mu_2 \tilde{Q}_{h_2}(1, \xi, \delta) = \mu_2 \tilde{Q}_{h_2-1}(1, \xi, \delta), \tag{3.18}$$

$$\mu_2 \tilde{Q}_{k_1+1}(1, \xi, \delta) = \mu_1 \tilde{Q}_{k_1}(1, \xi, \delta) + \mu_1 \sum_{h=1}^{k_1-1} \phi_{1,h}(\xi, \delta) + \mu_2 \phi_{0,k_1+k_2}(\xi, \delta), \tag{3.19}$$

$$\mu_1 \tilde{Q}_{h_1}(1, \xi, \delta) = \mu_1 \tilde{Q}_{h_1-1}(1, \xi, \delta) - \mu_1 \phi_{1,h_1-1}(\xi, \delta), \tag{3.20}$$

$$\mu_1 \tilde{Q}_1(1, \xi, \delta) = \mu_2 \tilde{Q}_{k_1+k_2}(1, \xi, \delta) - \mu_2 \phi_{0,k_1+k_2}(\xi, \delta), \tag{3.21}$$

for  $h_1 = 2, \dots, k_1$  and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . The summation of (3.18)–(3.21) over all  $h = 1, 2, \dots, k_1 + k_2$  cancels all terms. ■

**Proposition 3.4.** *After taking the summation over all  $n_1 = 0, 1, 2, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$  of the Taylor series of all perturbed balance equations, substituting (3.5) and using the results from Appendix A, the  $\mathcal{O}(\delta)$  terms cancel.*

PROOF: Define

$$\phi'_{n_1,h}(\xi, \delta) := \frac{\partial \phi_{n_1,h}(\xi, \delta)}{\partial \xi},$$

for  $n_1 = 0, 1, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$ . Given the fact that the  $\mathcal{O}(1)$  terms cancel, taking the  $\mathcal{O}(\delta)$  terms leads to the following equations:

$$\begin{aligned} 0 &= \mu_2 \phi'_{0,k_1+k_2}(\xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{0,k_1+1}(\xi, \delta), \\ 0 &= \mu_2 \phi'_{0,h_2-1}(\xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{0,h_2}(\xi, \delta), \\ 0 &= \mu_2 \phi'_{1,k_1+k_2}(\xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{1,1}(\xi, \delta), \\ 0 &= -\mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{1,h_1}(\xi, \delta), \\ 0 &= \mu_2 \phi'_{n_1+1,k_1+k_2}(\xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{n_1+1,1}(\xi, \delta), \\ 0 &= -\mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{n_1+1,h_1}(\xi, \delta), \\ 0 &= -\mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{n_1,k_1+1}(\xi, \delta), \\ 0 &= \mu_2 \phi'_{n_1,h_2-1}(\xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \phi'_{n_1,h_2}(\xi, \delta), \end{aligned}$$

for  $h_1 = 2, \dots, k_1$  and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . Taking the generating functions of these equations and substituting  $z = 1$  results in the following set of equations:

$$0 = \mu_2 \phi'_{0,k_1+k_2}(\xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_{k_1+1}(1, \xi, \delta), \tag{3.22}$$

$$0 = \mu_2 \tilde{Q}'_{h_2-1}(1, \xi, \delta) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_{h_2}(1, \xi, \delta), \tag{3.23}$$

$$0 = \mu_2 \left(\tilde{Q}'_{k_1+k_2}(1, \xi, \delta) - \phi'_{0,k_1+k_2}(\xi, \delta)\right) - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_1(1, \xi, \delta), \tag{3.24}$$

$$0 = -\mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_{h_1}(1, \xi, \delta), \tag{3.25}$$

for  $h_1 = 2, \dots, k_1$  and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . The PGF  $\tilde{Q}'_h(z, \xi, \delta)$  is the derivative of  $\tilde{Q}_h(z, \xi, \delta)$  with respect to  $\xi$ .

The summation of (3.22)–(3.25) over all  $h = 1, 2, \dots, k_1 + k_2$  yields:

$$\mu_2 \sum_{h_1=1}^{k_1} \tilde{Q}'_{h_1}(1, \xi, \delta) = \mu_2 \frac{\lambda_1}{\mu_1} \sum_{h=1}^{k_1+k_2} \tilde{Q}'_h(1, \xi, \delta). \tag{3.26}$$

Apparently, the  $\mathcal{O}(\delta)$  terms do not cancel (yet). However, after substituting (3.5), taking the  $\mathcal{O}(\delta)$  terms, and using (3.7), we obtain:

$$\mu_2 \sum_{h_1=1}^{k_1} \tilde{L}_{h_1}^{(0)}(1)P'_0(\xi) = \mu_2 \frac{\lambda_1}{\mu_1} \sum_{h=1}^{k_1+k_2} \tilde{L}_h^{(0)}(1)P'_0(\xi). \tag{3.27}$$

Since the (at this moment still unknown) terms  $P'_0(\xi)$  cancel out, and since  $\sum_{h=1}^{k_1+k_2} \tilde{L}_h^{(0)}(1) = 1$ , Eq. (3.27) reduces to

$$\sum_{h_1=1}^{k_1} \tilde{L}_{h_1}^{(0)}(1) = \rho_1,$$

which is indeed true (see (A.5)). ■

**Proposition 3.5.** *Taking the summation over all  $n_1 = 0, 1, 2, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$  of the Taylor series of all perturbed balance equations and equating the  $\mathcal{O}(\delta^2)$  terms, yields the following differential equation for  $P_0(\xi)$ :*

$$\omega P'_0(\xi) = - \left( \mu_2 + \frac{\lambda_1 \mu_2 (\mu_2 - \mu_1)}{\mu_1^2} \right) P''_0(\xi). \tag{3.28}$$

PROOF: Define

$$\phi''_{n_1,h}(\xi, \delta) := \frac{\partial^2 \phi_{n_1,h}^{(0)}(\xi, \delta)}{\partial \xi^2},$$

for  $n_1 = 0, 1, \dots$  and  $h = 1, 2, \dots, k_1 + k_2$ . As before, the  $\mathcal{O}(1)$  terms cancel. From the proof of the Proposition 3.4 we have learned to include  $\mathcal{O}(\delta)$  terms as well, because multiplied by  $\delta \phi'_{n,h}^{(1)}(\xi)$  these terms are  $\mathcal{O}(\delta^2)$  as well. This leads to the following equations:

$$\begin{aligned} 0 &= \mu_2 \left[ \phi'_{0,k_1+k_2}{}^{(1)}(\xi) + \frac{1}{2} \phi''_{0,k_1+k_2}{}^{(0)}(\xi) \right] - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{0,k_1+1}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{0,k_1+1}{}^{(0)}(\xi) \right] \\ &\quad + \omega \phi'_{0,k_1+1}{}^{(0)}(\xi, \delta), \\ 0 &= \mu_2 \left[ \phi'_{0,h_2-1}{}^{(1)}(\xi) + \frac{1}{2} \phi''_{0,h_2-1}{}^{(0)}(\xi) \right] - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{0,h_2}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{0,h_2}{}^{(0)}(\xi) \right] + \omega \phi'_{0,h_2}{}^{(0)}(\xi), \\ 0 &= \mu_2 \left[ \phi'_{1,k_1+k_2}{}^{(1)}(\xi) + \frac{1}{2} \phi''_{1,k_1+k_2}{}^{(0)}(\xi) \right] - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{1,1}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{1,1}{}^{(0)}(\xi) \right] + \omega \phi'_{1,1}{}^{(0)}(\xi), \\ 0 &= -\mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{1,h_1}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{1,h_1}{}^{(0)}(\xi) \right] + \omega \phi'_{1,h_1}{}^{(0)}(\xi), \\ 0 &= \mu_2 \left[ \phi'_{n_1+1,k_1+k_2}{}^{(1)}(\xi) + \frac{1}{2} \phi''_{n_1+1,k_1+k_2}{}^{(0)}(\xi) \right] - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{n_1+1,1}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{n_1+1,1}{}^{(0)}(\xi) \right] \\ &\quad + \omega \phi'_{n_1+1,1}{}^{(0)}(\xi), \\ 0 &= -\mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{n_1+1,h_1}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{n_1+1,h_1}{}^{(0)}(\xi) \right] + \omega \phi'_{n_1+1,h_1}{}^{(0)}(\xi), \\ 0 &= -\mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{n_1,k_1+1}{}^{(1)}(\xi) - \frac{1}{2} \phi''_{n_1,k_1+1}{}^{(0)}(\xi) \right] + \omega \phi'_{n_1,k_1+1}{}^{(0)}(\xi), \end{aligned}$$

$$0 = \mu_2 \left[ \phi'_{n_1, h_2-1}(1, \xi) + \frac{1}{2} \phi''_{n_1, h_2-1}(1, \xi) \right] - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ \phi'_{n_1, h_2}(1, \xi) - \frac{1}{2} \phi''_{n_1, h_2}(1, \xi) \right] + \omega \phi'_{n_1, h_2}(1, \xi),$$

for  $h_1 = 2, \dots, k_1$  and  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . As we have done already a couple of times before, we can use the generating functions to easily sum all of these equations. Each equation contains the following three types of terms:  $\phi'_{n_1, h}(1, \xi)$ ,  $\phi''_{n_1, h}(1, \xi)$ , and  $\omega \phi'_{n_1, h}(1, \xi)$ . We denote the corresponding generating functions with  $\tilde{Q}'_h(z, \xi)$ ,  $\tilde{Q}''_h(z, \xi)$ , and  $\tilde{Q}^{(0)}_h(z, \xi)$ . After summing all equations, we obtain the following equation:

$$\mu_2 \frac{\lambda_1}{\mu_1} \sum_{h=1}^{k_1+k_2} \tilde{Q}'_h(1, \xi) - \mu_2 \sum_{h_1=1}^{k_1} \tilde{Q}'_{h_1}(1, \xi) \tag{3.29a}$$

$$+ \frac{1}{2} \mu_2 \sum_{h_2=k_1+1}^{k_1+k_2} \tilde{Q}''_{h_2}(1, \xi) + \frac{1}{2} \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \sum_{h=1}^{k_1+k_2} \tilde{Q}''_h(1, \xi) \tag{3.29b}$$

$$+ \omega \sum_{h=1}^{k_1+k_2} \tilde{Q}^{(0)}_h(1, \xi) \tag{3.29c}$$

$$= 0.$$

Note that the derivation of (3.29a) and (3.29b) follows closely the manner in which (3.26) has been derived. The last term (3.29c) follows directly from collecting the  $\omega$ -terms. Using the results obtained in appendices the equation can be rewritten to

$$\frac{\lambda_1 \mu_2^2}{\mu_1^2} P_0''(\xi) + \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) P_0''(\xi) + \omega P_0'(\xi) = 0. \tag{3.30}$$

The proof that (3.29a) is the same as the first term in (3.30) can be found in Appendix A. The second and third term follow from (A.5) and the fact that  $\tilde{Q}_h^{(0)}(1, \xi) = \tilde{L}_h^{(0)}(1) P_0(\xi)$ . Equation (3.30) can be rewritten to (3.28), which concludes this proof. ■

### 3.6. The Scaled Number of Customers in the Critically Loaded Queue

Now we can finally present the density of the scaled number of customers in  $Q_2$ , denoted by  $P_0(\xi)$ . It is obtained by solving the differential equation (3.28):

$$P_0(\xi) = \eta e^{-\eta \xi}, \tag{3.31}$$

with

$$\eta = \omega \left[ \mu_2 + \frac{\lambda_1 \mu_2 (\mu_2 - \mu_1)}{\mu_1^2} \right]^{-1}.$$

We have used that  $\int_0^\infty P_0(\xi) d\xi = 1$  and that  $\sum_{n_1=0}^\infty \sum_{n_2=0}^\infty \sum_{h=1}^{k_1+k_2} p(n_1, n_2, h) = 1$ . Without loss of generality, we may take  $\omega = \mu_2$ , which means that

$$\frac{1}{\eta} = 1 - \frac{\lambda_1}{\mu_1} + \mu_2 \frac{\lambda_1}{\mu_1^2}. \tag{3.32}$$

We motivate this choice for  $\omega$  by noting that we consider the scaled queue length  $\delta N_2$  for  $\delta \downarrow 0$ . By choosing  $\omega = \mu_2$ , and using (3.2), our scaling becomes equivalent to considering the

scaled queue length  $(1 - \rho)N_2$ , which is commonly used. Finally, by applying the multiclass distributional law of Bertsimas and Mourtzinou [2] it directly follows that the scaled waiting time at  $Q_2$  also follows an exponential distribution with parameter  $\lambda_2\eta$ .

### 3.7. Main Result

The analysis of the present section has the following immediate consequence for the joint (scaled) queue-length distribution in heavy traffic, which is the main result of this paper.

*Main result* For  $\lambda_1/k_1 < \lambda_2/k_2$  and  $\lambda_2 = \mu_2((1 - \rho_1) - \delta)$ , we have:

$$\lim_{\delta \downarrow 0} \mathbb{P}[N_1 \leq n_1, \delta N_2 \leq \xi] = \mathcal{L}(n_1) (1 - e^{-\eta\xi}), \quad (3.33)$$

where  $\mathcal{L}(\cdot)$  is the cumulative probability distribution of the queue length of a queueing system with multiple vacations satisfying Properties P3.3–P3.3, and  $\eta$  is given by (3.32).

## 4. FINAL REMARKS AND SUGGESTIONS FOR FURTHER RESEARCH

*Interpretation* The main result (3.33) derived in the preceding section has the following intuitively appealing interpretation:

1. The number of customers in the stable queue has the same distribution as the number of customers in a  $k$ -limited vacation system with Erlang- $k_2$  distributed vacations.
2. The scaled number of customers in the critically loaded queue is exponentially distributed with parameter  $\eta$ .
3. The number of customers in the stable queue and the (scaled) number of customers in the critically loaded queue are independent.

Below we explain these properties heuristically.

Property 1 can be explained by the fact that if  $Q_2$  is in heavy traffic, then exactly  $k_2$  customers are served at this queue during each cycle. If we place an outside observer at  $Q_1$ , then, from his perspective, this queue behaves like a  $k$ -limited vacation model in heavy traffic, where the vacation distribution is given by the convolutions of  $k_2$  exponentially distributed service times distributions in  $Q_2$ .

For Property 2, we note that the total workload in the system equals the amount of work in an  $M/G/1$  queue in which the two customer classes are combined into one customer class with arrival rate  $\lambda_1 + \lambda_2$  and hyperexponentially distributed service times, that is, the service time is with probability  $\lambda_i/(\lambda_1 + \lambda_2)$  exponentially distributed with parameter  $\mu_i$ ,  $i = 1, 2$ . Based on standard heavy-traffic results for the  $M/G/1$  queue, this implies that the distribution of the scaled total workload converges to an exponential distribution with mean  $\rho\mathbb{E}[R]$ , where  $R$  is a residual service time. For a hyperexponential distribution, we have

$$\mathbb{E}[R] = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{\rho},$$

which implies that the total (scaled) asymptotic workload is exponentially distributed with parameter  $\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2$ . In heavy traffic, almost all customers are located in  $Q_2$  so the total number of customers at this queue is exponentially distributed with mean  $\mu_2(\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2)$ . Since  $\lambda_2 \uparrow \mu_2(1 - \lambda_1/\mu_1)$ , the scaled number of customers in  $Q_2$  is exponentially distributed with parameter  $\eta$ . Using the multiclass distributional law of Bertsimas and

Mourtzinou [2], it can be shown that the scaled asymptotic waiting time of customers in  $Q_2$  is exponentially distributed with parameter  $\lambda_2\eta$ .

Finally, Property 3 follows from the time-scale separation in heavy traffic which implies that the dynamics of the stable queue evolve at a much faster time scale than the dynamics of the critically loaded queue. Since the amount of “memory” of the stable queue asymptotically vanishes compared to that of the critically loaded queue, the number of customers in  $Q_1$  and the scaled queue length of  $Q_2$  become independent in the limit.

*Two critically loaded queues* In the present paper, we have analyzed the heavy-traffic behavior in case only  $Q_2$  becomes critically loaded, that is, when Assumption (2.1) is satisfied. The limiting regime in which both queues become saturated simultaneously ( $\lambda_1/k_1 = \lambda_2/k_2$ ), shows fundamentally different system behavior. That is, for general  $\rho$  the waiting time at  $Q_1$  is an (unknown) function of the visit time at  $Q_2$  and  $1/(1 - \rho)$ . This implies that  $\mathcal{O}(1 - \rho)$  variations in the visit time at  $Q_2$  are relevant for the heavy-traffic behavior at  $Q_1$ . More colloquially, it is not sufficient anymore to use a scaling that implies that exactly  $k_2$  customers are served at  $Q_2$  during each cycle, that is, the probability that there are served less than  $k_2$  customers cannot be neglected, when analyzing the asymptotic behavior of  $Q_1$ .

*Further research* The analysis in this paper allows different kind of extensions. First, one could consider phase-type interarrival-time or service-time distributions. The approach introduced in the present paper may be extended, without fundamentally changing the analysis, to such a system. Another extension could be the introduction of switchover times whenever the server switches between queues. Such an extension requires more severe adaptations to the approach and the analysis, and is the topic of a forthcoming paper. Finally, we want to mention that the singular-perturbation technique can also be applied to derive the heavy-traffic analysis of a system consisting of more than two (say  $N$ ) queues, with one queue becoming critically loaded. Following the lines of the current paper, one can show that in this limiting regime the stable queues have the same joint queue-length distribution as in a polling model with  $N - 1$  queues and an extended switchover time, whereas the scaled queue-length distribution of the critically loaded queue is again exponentially distributed. As such, the results of the present paper provide a theoretical basis for the transformation of large polling systems into smaller systems for approximation purposes; cf. LaPadula and Levy [20].

#### Acknowledgement

The authors would like to thank Sem Borst and Onno Boxma for interesting discussions and valuable comments on earlier drafts of this paper.

#### References

1. Adan, I.J.B.F., van Leeuwen, J.S.H. & Winands, E.M.M. (2006). On the application of Rouché's theorem in queueing theory. *Operations Research Letters* 34(3): 355–360.
2. Bertsimas, D. & Mourtzinou, G. (1997). Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws. *Operations Research* 45(3): 470–487.
3. Blanc, J.P.C. (1992). An algorithmic solution of polling models with limited service disciplines. *IEEE Transactions of Communications* 40(7): 1152–1155.
4. Boon, M.A.A., Adan, I.J.B.F., Winands, E.M.M. & Down, D.G. (2012). Delays at signalised intersections with exhaustive traffic control. *Probability in the Engineering and Informational Sciences* 26(3): 337–373.

5. Borst, S.C., Boxma, O.J. & Levy, H. (1995). The use of service limits for efficient operation of multistation single-medium communication systems. *IEEE/ACM Transactions on Networking* 3(5): 602–612.
6. Boxma, O.J. (1985). Two symmetric queues with alternating service and switching times. In E. Gelenbe (ed), *Performance '84*, North-Holland, Amsterdam, pp. 409–431.
7. Boxma, O.J. & Groenendijk, W.P. (1988). Two queues with alternating service and switching times. In O.J. Boxma & R. Syski (eds), *Queueing Theory and its Applications — Liber Amicorum for J.W. Cohen*, North-Holland, Amsterdam, pp. 261–282.
8. Charzinski, J., Renger, T. & Tangemann, M. (1994). Simulative comparison of the waiting time distributions in cyclic polling systems with different service strategies. In *Proceedings of the 14th International Teletraffic Congress, Antibes Juan-les-Pins*, pp. 719–728.
9. Coffman, Jr., E.G., Puhalskii, A.A. & Reiman, M.I. (1995). Polling systems with zero switchover times: a heavy-traffic averaging principle. *The Annals of Applied Probability* 5(3): 681–719.
10. Coffman, Jr., E.G., Puhalskii, A.A. & Reiman, M.I. (1998). Polling systems in heavy-traffic: a Bessel process limit. *Mathematics of Operations Research* 23: 257–304.
11. Cohen, J.W. & Boxma, O.J. (1981). The  $M/G/1$  queue with alternating service formulated as a Riemann-Hilbert problem. In F.J. Kylstra (ed), *Performance '81*, pp. 181–189.
12. Down, D.G. (1998). On the stability of polling models with multiple servers. *Journal of Applied Probability*, 35: 925–935.
13. Eisenberg, M. (1979). Two queues with alternating service. *SIAM Journal on Applied Mathematics* 36(2): 287–303.
14. Fricker, C. & Jaibi, M.R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* 15: 211–238.
15. Fuhrmann, S.W. (1981). Performance analysis of a class of cyclic schedules. Technical Memorandum 81-59531-1, Bell Laboratories.
16. Groenendijk, W.P. (1990). *Conservation Laws in Polling Systems*. PhD thesis, University of Utrecht.
17. Ibe, O.C. (1990). Analysis of polling systems with mixed service disciplines. *Stochastic Models* 6: 667–689.
18. Keilson, J. & Servi, L.D. (1990). The distributional form of Little's Law and the Fuhrmann–Cooper decomposition. *Operations Research Letters* 9(4): 239–247.
19. Knessl, C. & Tier, C. (1995). Applications of singular perturbation methods in queueing. In *Advances in Queueing Theory, Methods, and Open Problems, Probability and Stochastics Series*. Boca Raton: CRC Press, pp. 311–336.
20. LaPadula, C.A. & Levy, H. (1996). Customer delay in very large multi-queue single-server systems. *Performance Evaluation* 26(3): 201–218.
21. Lee, D.-S. (1996). A two-queue model with exhaustive and limited service disciplines. *Stochastic Models* 12(2): 285–305.
22. Lee, T.T. (1989).  $M/G/1/N$  queue with vacation time and limited service discipline. *Performance Evaluation* 9(3): 181–190.
23. Morrison, J.A. & Borst, S.C. (2010). Interacting queues in heavy traffic. *Queueing Systems* 65(2): 135–156.
24. Ozawa, T. (1990). Alternating service queues with mixed exhaustive and  $K$ -limited services. *Performance Evaluation*, 11: 165–175.
25. Ozawa, T. (1997). Waiting time distribution in a two-queue model with mixed exhaustive and gated-type  $K$ -limited services. In *Proceedings of International Conference on the Performance and Management of Complex Communication Networks*, pp. 231–250.
26. Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* 13: 409–426.
27. Van der Mei, R.D. (2007). Towards a unifying theory on branching-type polling models in heavy traffic. *Queueing Systems* 57: 29–46.
28. Winands, E.M.M., Adan, I.J.B.F., van Houtum, G.J. & Down, D.G. (2009). A state-dependent polling model with  $k$ -limited service. *Probability in the Engineering and Informational Sciences* 23(2): 385–408.

## APPENDIX

### Appendix A. A Vacation Model with $k$ -limited Service

In this appendix, we study a queueing model with multiple vacations and  $k$ -limited service. The main goal is to find the PGF of the queue-length distribution, as to prove (3.9)–(3.11). At the end

of this appendix some additional properties of this queue-length distribution are given, which will be used in Section 3.5 and in Appendix 4.

The service times in this vacation model are exponentially distributed with parameter  $\mu_1$ , and the vacation length is Erlang( $k_2$ ) distributed with parameter  $\mu_2$ . The service discipline is  $k$ -limited service, with at most  $k_1$  customers being served during one visit period. Although the queue-length distribution for the case with generally distributed service and vacation times has been studied by Lee [22], we provide the proof here to keep the paper self-contained, but also because our state space is slightly different and we do not look at embedded epochs, yielding slightly different expressions than in [22].

The starting point is to obtain generating functions from the balance Eqs. (3.6a)–(3.6h). Multiplying Eq. (3.6h) with  $z^{n_1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.6b), yields the following equation:

$$\tilde{L}_{h_2}^{(0)}(z) = \tilde{H}(z)\tilde{L}_{h_2-1}^{(0)}(z), \tag{A.1}$$

for  $h_2 = k_1 + 2, \dots, k_1 + k_2$ , where  $\tilde{H}(z)$  is defined in (3.8). The interpretation of (A.1) is that the number of customers in the system during a certain vacation stage is simply the number of customers present at the previous stage of the vacation, plus the arrivals during one (exponentially distributed) stage. Obviously, no customers leave the system during a vacation.

Multiplying Eq. (3.6g) with  $z^{n_1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.6a), yields the following equation:

$$\tilde{L}_{k_1+1}^{(0)}(z) = \tilde{H}(z) \left( \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} \pi_{1,h}^{(0)} + \pi_{0,k_1+k_2}^{(0)} + \frac{\mu_1}{\mu_2 z} \tilde{L}_{k_1}^{(0)}(z) \right). \tag{A.2}$$

Multiplying Eq. (3.6f) with  $z^{n_1+1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.6d) multiplied by  $z$ , yields the following equation:

$$\tilde{L}_{h_1}^{(0)}(z) = \frac{\tilde{G}(z)}{z} \left( \tilde{L}_{h_1-1}^{(0)}(z) - \pi_{1,h_1-1}^{(0)} \right), \tag{A.3}$$

for  $h_1 = 2, 3, \dots, k_1$ , where  $\tilde{G}(z)$  is defined in (3.8).

Multiplying Eq. (3.6e) with  $z^{n_1+1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.6c) multiplied by  $z$ , yields the following equation:

$$\tilde{L}_1^{(0)}(z) = \tilde{G}(z) \frac{\mu_2}{\mu_1} \left( \tilde{L}_{k_1+k_2}^{(0)}(z) - \pi_{0,k_1+k_2}^{(0)} \right). \tag{A.4}$$

We now have  $k_1 + k_2$  equations, each of which expresses  $\tilde{L}_h^{(0)}(z)$  in terms of  $\tilde{L}_{h-1}^{(0)}(z)$  (and  $\tilde{L}_1^{(0)}(z)$  in terms of  $\tilde{L}_{k_1+k_2}^{(0)}(z)$ ). Finally, we can solve these equations to find the expressions for  $\tilde{L}_h^{(0)}(z)$ , for  $h = 1, 2, \dots, k_1 + k_2$ . The results are given in (3.9)–(3.11).

Note that there are still  $k_1$  unknowns:  $\pi_{1,h_1}^{(0)}$ , for  $h_1 = 1, \dots, k_1 - 1$ , and  $\pi_{0,k_1+k_2}^{(0)}$ . These can be found using the roots of the denominator of (3.9). Rouché’s Theorem implies that the denominator has  $k_1$  roots on and inside the unit circle. The requirement that (3.9) should be regular inside the unit circle, implies that the numerator of (3.9) should have these same roots [1]. Hence, we have a set of equations involving the roots and the numerator of (3.9) to eliminate these  $k_1$  unknowns.

*Some additional properties* In this paragraph, we derive some results that are used throughout this paper, particularly in Section 3.5 and in Appendix 4. From a balancing argument, we know that the fraction of time that the system is in a vacation is  $1 - \rho$  (where, in this system,  $\rho = \lambda_1/\mu_1$ ).



Conversely, the fraction of time that the system is serving customers is  $\rho$ . Hence,

$$\sum_{h=1}^{k_1} \tilde{L}_h^{(0)}(1) = \frac{\lambda_1}{\mu_1}, \quad \sum_{h=k_1+1}^{k_1+k_2} \tilde{L}_h^{(0)}(1) = k_2 \tilde{L}_{k_1+k_2}^{(0)}(1) = 1 - \frac{\lambda_1}{\mu_1}. \tag{A.5}$$

Moreover, from (A.1) we know that  $\tilde{L}_{h_2}^{(0)}(1) = \tilde{L}_{h_2-1}^{(0)}(1)$  for  $h_2 = k_1 + 2, \dots, k_1 + k_2$ . It follows that

$$\tilde{L}_{h_2}^{(0)}(1) = \frac{1}{k_2} \left( 1 - \frac{\lambda_1}{\mu_1} \right), \quad h_2 = k_1 + 1, \dots, k_1 + k_2.$$

From (A.3) and (A.4) we now have

$$\tilde{L}_{h_1}^{(0)}(1) = \frac{\mu_2}{\mu_1} \left( \frac{1}{k_2} \left( 1 - \frac{\lambda_1}{\mu_1} \right) - \pi_{0,k_1+k_2}^{(0)} \right) - \sum_{i=1}^{h_1-1} \pi_{1,i}^{(0)}, \quad h_1 = 1, \dots, k_1. \tag{A.6}$$

The following relation for the unknowns  $\pi_{0,k_1+k_2}^{(0)}$  and  $\pi_{1,h_1}^{(0)}$  ( $h_1 = 1, \dots, k_1 - 1$ ) can be derived by combining all of these results:

$$k_1 \pi_{0,k_1+k_2}^{(0)} + \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} (k_1 - h) \pi_{1,h}^{(0)} = \frac{1}{k_2} \left[ k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) - k_2 \frac{\lambda_1}{\mu_2} \right]. \tag{A.7}$$

This relation turns out to be crucial to derive many results in this paper, without having to know the exact expressions for all of the individual probabilities.

**Remark A.1.** A balancing argument has been the starting point to derive all of the above properties. A rigorous way to derive these results, is by using L'Hôpital's rule on (3.9) to determine  $\tilde{L}_{k_1+k_2}^{(0)}(1)$ , and subsequently deriving an expression for  $\sum_{h=1}^{k_1+k_2} \tilde{L}_h^{(0)}(1)$ , which we know is equal to one.

**Appendix B. The second-order perturbed balance equations**

The main goal of this appendix is to prove that (3.29a) can be written as the first term in (3.30). If we rearrange the summations slightly, we can write the equation that we need to prove as follows:

$$\mu_2 \frac{\lambda_1}{\mu_1} \sum_{h_2=k_1+1}^{k_1+k_2} \tilde{Q}'_{h_2}(1, \xi) - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \sum_{h_1=1}^{k_1} \tilde{Q}'_{h_1}(1, \xi) = \frac{\lambda_1 \mu_2^2}{\mu_1^2} P_0''(\xi). \tag{B.1}$$

This equation should follow from Eqs. (3.16a)–(3.16h). In order to prove it, we take the following steps:

1. First, we take the generating functions of Eqs. (3.16a)–(3.16h) to develop relations for  $\tilde{Q}'_h(1)(z, \xi)$  ( $h = 1, \dots, k_1 + k_2$ ).
2. The next step involves solving these equations to find an expression for  $\tilde{Q}'_{k_1+k_2}(1)(z, \xi)$ .
3. Step 3 is to reformulate (B.1) in terms of  $\tilde{Q}'_{k_1+k_2}(1)(\xi)$ . It turns out that in this stage all terms containing probabilities  $\phi'_{n,h}(1)(\xi)$  (or their generating functions) are eliminated.
4. The last step involves some more algebraic manipulations which eliminate all terms containing probabilities  $\phi''_{n,h}(0)(\xi)$  and, eventually, prove (B.1).

*Step 1: Find relations for the generating functions* Multiplying Eq. (3.16b) with  $z^{n_1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.16b), yields the following equation:

$$\tilde{Q}_{h_2}^{(1)}(z, \xi) = \tilde{H}(z) \left( \tilde{Q}_{h_2-1}^{(1)}(z, \xi) + \tilde{Q}'_{h_2-1}(0)(z, \xi) - \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_{h_2}(0)(z, \xi) \right), \tag{B.2}$$

for  $h_2 = k_1 + 2, \dots, k_1 + k_2$ , where  $\tilde{H}(z)$  is defined in (3.8).

Multiplying Eq. (3.16g) with  $z^{n_1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.16a), yields the following equation:

$$\begin{aligned} \tilde{Q}_{k_1+1}^{(1)}(z, \xi) = \tilde{H}(z) & \left( \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} \phi_{1,h}^{(1)}(\xi) + \phi_{0,k_1+k_2}^{(1)}(\xi) + \frac{\mu_1}{z\mu_2} \tilde{Q}_{k_1}^{(1)}(z, \xi) + \phi'_{0,k_1+k_2}(0)(\xi) \right. \\ & \left. - \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_{k_1+1}(0)(z, \xi) \right). \end{aligned} \tag{B.3}$$

Multiplying Eq. (3.16f) with  $z^{n_1+1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.16d) multiplied by  $z$ , yields the following equation:

$$\tilde{Q}_{h_1}^{(1)}(z, \xi) = \frac{\tilde{G}(z)}{z} \left( \tilde{Q}_{h_1-1}^{(1)}(z, \xi) - \phi_{1,h_1-1}^{(1)}(\xi)z - \frac{z\mu_2}{\mu_1} \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_{h_1}(0)(z, \xi) \right), \tag{B.4}$$

for  $h_1 = 2, 3, \dots, k_1$ , where  $\tilde{G}(z)$  is defined in (3.8).

Multiplying Eq. (3.16e) with  $z^{n_1+1}$ , summing over all  $n_1 = 1, 2, \dots$ , and adding Eq. (3.16c) multiplied by  $z$ , yields the following equation:

$$\begin{aligned} \tilde{Q}_1^{(1)}(z, \xi) = \frac{\mu_2}{\mu_1} \tilde{G}(z) & \left( \tilde{Q}_{k_1+k_2}^{(1)}(z, \xi) - \phi_{0,k_1+k_2}^{(1)}(\xi) + \tilde{Q}'_{k_1+k_2}(0)(z, \xi) - \phi'_{0,k_1+k_2}(0)(\xi) \right. \\ & \left. - \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{Q}'_1(0)(z, \xi) \right). \end{aligned} \tag{B.5}$$

*Step 2: Solve these relations and determine  $\tilde{Q}'_{k_1+k_2}(1)(z, \xi)$*  Solving Eqs. (B.2)–(B.5) requires a lot of straightforward, but tedious, computations. For reasons of compactness we will present some relevant intermediate results in this appendix, but leave the exact derivations to the reader. Firstly, one can use (B.2), combined with (A.1), to express  $\tilde{Q}_{k_1+k_2}^{(1)}(z, \xi)$  in terms of  $\tilde{Q}_{k_1+1}^{(1)}(z, \xi)$  and  $\tilde{Q}'_{k_1+k_2}(0)(z, \xi)$ :

$$\tilde{Q}_{k_1+k_2}^{(1)}(z, \xi) = \tilde{H}(z)^{k_2-1} \tilde{Q}_{k_1+1}^{(1)}(z, \xi) + (k_2 - 1) \left(1 - \left(1 - \frac{\lambda_1}{\mu_1}\right) \tilde{H}(z)\right) \tilde{Q}'_{k_1+k_2}(0)(z, \xi). \tag{B.6}$$

Second, we use (B.3) to express  $\tilde{Q}_{k_1+1}^{(1)}(z, \xi)$  in terms of  $\tilde{Q}_{k_1}^{(1)}(z, \xi)$  and  $\tilde{Q}'_{k_1+1}(0)(z, \xi)$ . Subsequently, we use (B.4) and (A.3) to express  $\tilde{Q}_{k_1}^{(1)}(z, \xi)$  in terms of  $\tilde{Q}_1^{(1)}(z, \xi)$  and  $\tilde{Q}'_h(0)(z, \xi)$ :

$$\begin{aligned} \tilde{Q}_{k_1}^{(1)}(z, \xi) = & \left(\frac{\tilde{G}(z)}{z}\right)^{k_1-1} \tilde{Q}_1^{(1)}(z, \xi) - \sum_{h=1}^{k_1-1} \phi_{1,h}^{(1)}(\xi)z \left(\frac{\tilde{G}(z)}{z}\right)^{k_1-h} \\ & - \frac{z\mu_2}{\mu_1} \left(1 - \frac{\lambda_1}{\mu_1}\right) \sum_{h=2}^{k_1} \left(\frac{\tilde{G}(z)}{z}\right)^{k_1-h+1} \tilde{Q}'_h(0)(z, \xi). \end{aligned} \tag{B.7}$$

Finally, we use (B.5) to express  $\tilde{Q}_1^{(1)}(z, \xi)$  in terms of  $\tilde{Q}_{k_1+k_2}^{(1)}(z, \xi)$  again. After some rearrangement of the terms, this leads to the following expression for  $\tilde{Q}_{k_1+k_2}^{(1)}(z, \xi)$ :

$$\tilde{Q}_{k_1+k_2}^{(1)}(z, \xi) = \frac{A^{(1)}(z, \xi) + A'^{(0)}(z, \xi)}{D(z)}, \tag{B.8}$$

where

$$A^{(1)}(z, \xi) = \tilde{H}(z)^{k_2} \left[ \phi_{0, k_1+k_2}^{(1)}(\xi) \left( 1 - \left( \tilde{G}(z)/z \right)^{k_1} \right) + \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} \phi_{1, h}^{(1)}(\xi) \left( 1 - \left( \tilde{G}(z)/z \right)^{k_1-h} \right) \right], \tag{B.9}$$

$$A'^{(0)}(z, \xi) = k_2 \tilde{Q}'_{k_1+k_2}{}^{(0)}(z, \xi) - \tilde{H}(z)^{k_2} \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} \phi'_{1, h}{}^{(0)}(\xi) \left( 1 - \left( \tilde{G}(z)/z \right)^{k_1-h} \right) - \left( 1 - \frac{\lambda_1}{\mu_2} \right) \left[ \tilde{H}(z)^{k_2} \sum_{h=1}^{k_1} \tilde{Q}'_h{}^{(0)}(z, \xi) \left( \tilde{G}(z)/z \right)^{k_1-h+1} + k_2 \tilde{H}(z) \tilde{Q}'_{k_1+k_2}{}^{(0)}(z, \xi) \right], \tag{B.10}$$

$$D(z) = 1 - \left( \tilde{G}(z)/z \right)^{k_1} \tilde{H}(z)^{k_2}. \tag{B.11}$$

Note that  $A^{(1)}(z, \xi)$  only contains the probabilities  $\phi_{n, h}^{(1)}(\xi)$ . All probabilities  $\phi'_{0, k_1+k_2}(\xi)$ , and their generating functions, are contained in  $A'^{(0)}(z, \xi)$ . Also note that  $\frac{A^{(1)}(z, \xi)}{D(z)}$  is exactly the same expression as (3.9), but with constants  $\pi_{n, h}^{(0)}$  replaced by  $\phi_{n, h}^{(1)}$ . The reason is that, if one would ignore the probabilities  $\phi'_{n, h}{}^{(0)}(\xi)$  in the balance Eqs. (3.16a)–(3.16h), the system is completely equivalent to the system (3.6a)–(3.6h), which corresponds to the vacation system studied in Appendix A.

*Step 3: Reformulate the original problem in terms of  $\tilde{Q}'_{k_1+k_2}{}^{(1)}(1, \xi)$*  In order to solve (B.1), we need to determine  $\sum_{h=1}^{k_1} \tilde{Q}'_{h_1}{}^{(1)}(1, \xi)$  and  $\sum_{h_1=k_1+1}^{k_1+k_2} \tilde{Q}'_{h_1}{}^{(1)}(1, \xi)$ . After substituting  $z = 1$  into Eqs. (B.2)–(B.5), we can express these sums in terms of  $\tilde{Q}'_{k_1+k_2}{}^{(1)}(1, \xi)$  and  $\tilde{Q}''_h{}^{(0)}(1, \xi)$ :

$$\sum_{h_1=1}^{k_1} \tilde{Q}'_{h_1}{}^{(1)}(1, \xi) = k_1 \frac{\mu_2}{\mu_1} \left( \tilde{Q}'_{k_1+k_2}{}^{(1)}(1, \xi) - \phi'_{0, k_1+k_2}{}^{(1)} \right) - \sum_{h=1}^{k_1-1} (k_1 - h) \phi'_{1, h}{}^{(1)} + k_1 \tilde{Q}''_1{}^{(0)}(1, \xi) - \frac{\mu_2}{\mu_1} \left( 1 - \frac{\lambda_1}{\mu_1} \right) \sum_{h=1}^{k_1} (k_1 - h + 1) \tilde{Q}''_h{}^{(0)}(1, \xi), \tag{B.12}$$

$$\sum_{h_1=k_1+1}^{k_1+k_2} \tilde{Q}'_{h_1}{}^{(1)}(1, \xi) = k_2 \tilde{Q}'_{k_1+k_2}{}^{(1)}(1, \xi) - \frac{\lambda_1}{\mu_1} \left( \frac{k_2 - 1}{2} \right) \left( 1 - \frac{\lambda_1}{\mu_1} \right) P_0''(\xi). \tag{B.13}$$

Since  $\tilde{Q}''_h{}^{(0)}(1, \xi)$  for  $h = 1, \dots, k_1$  can be determined directly using (A.6), it only remains to determine  $\tilde{Q}'_{k_1+k_2}{}^{(1)}(1, \xi)$ . Fortunately, according to the following lemma we can focus on the part  $\lim_{z \rightarrow 1} \frac{A'^{(0)}(z, \xi)}{D(z)}$  only.

**Lemma B.1.** *The probabilities  $\phi'_{0, k_1+k_2}{}^{(1)}$  and  $\phi'_{0, h}{}^{(1)}$  ( $h = 1, \dots, k_1 - 1$ ) in the left-hand side of Eq. (B.1) cancel out. Using (B.12) and (B.13), we can express this statement in a more formal presentation:*

$$\lim_{z \rightarrow 1} \mu_2 \frac{\lambda_1}{\mu_1} \left( k_2 \frac{A^{(1)}(z, \xi)}{D(z)} \right) - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left( k_1 \frac{\mu_2}{\mu_1} \left( \frac{A^{(1)}(z, \xi)}{D(z)} - \phi_{0, k_1+k_2}^{(1)} \right) - \sum_{h=1}^{k_1-1} (k_1 - h) \phi_{1, h}^{(1)} \right) = 0. \tag{B.14}$$

PROOF: Using L'Hôpital's rule it can be shown that

$$\lim_{z \rightarrow 1} \frac{A^{(1)}(z, \xi)}{D(z)} = \frac{\left(1 - \frac{\lambda_1}{\mu_1}\right) C^{(1)}}{k_1 \left(1 - \frac{\lambda_1}{\mu_1}\right) - k_2 \frac{\lambda_1}{\mu_2}}, \tag{B.15}$$

where

$$C^{(1)} = k_1 \phi_{0, k_1+k_2}^{(1)} + \frac{\mu_1}{\mu_2} \sum_{h=1}^{k_1-1} (k_1 - h) \phi_{1, h}^{(1)}.$$

After substituting (B.15) into (B.14), it is easily shown that the left-hand side of (B.14) indeed equals zero. ■

*Step 4: solve the original problem* We are almost ready to solve Eq. (B.1), but we compute two helpful intermediate results first.

**Lemma B.2.** Define  $Y := \lim_{z \rightarrow 1} \frac{\partial}{\partial z} \tilde{Q}'_{k_1+k_2}(z, \xi)^{(0)}$ . This can be written as

$$Y = \left[ 2 \frac{\mu_2}{\mu_1} \left( k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) - k_2 \frac{\lambda_1}{\mu_2} \right) \right]^{-1} \times \left[ \left( 1 - \frac{\lambda_1}{\mu_1} \right)^2 \sum_{h=1}^{k_1-1} h(k_1 - h) \phi'_{1, h}{}^{(0)} + \left( \frac{\lambda_1}{\mu_1} \left( 1 - \frac{\lambda_1}{\mu_1} \right) \right)^2 (k_1 + 1) - \frac{\lambda_1^2}{\mu_1 \mu_2} \left( 1 - \frac{\lambda_1}{\mu_1} \right) (k_2 - 1) + 2 \left( \frac{\lambda_1}{\mu_1} \right)^2 \right] P'_0(\xi). \tag{B.16}$$

PROOF: Equation (B.16) follows after differentiating (3.9) with respect to  $z$ , and subsequently applying L'Hôpital's rule twice. Obviously, some basic algebraic manipulations are required to obtain the presentation in (B.16). ■

**Lemma B.3.** Define  $X := \lim_{z \rightarrow 1} \frac{A'^{(0)}(z, \xi)}{D(z)}$ . This can be written as

$$X = -\frac{\mu_2}{\mu_1} Y + \left[ k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) - k_2 \frac{\lambda_1}{\mu_2} \right]^{-1} \times \left\{ \phi'_{0, k_1+k_2}{}^{(0)} \left( k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left( k_2 \frac{\lambda_1}{\mu_1} + \left( \frac{\lambda_1}{\mu_1} - k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \right) \frac{\mu_2}{\mu_1} \right) + \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left( k_2 \frac{\lambda_1}{\mu_2} + \frac{\lambda_1}{\mu_1} - \frac{\mu_1}{\mu_2} - k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \right) \sum_{h=1}^{k_1-1} (k_1 - h) \phi'_{1, h}{}^{(0)} + \left( 1 - \frac{\lambda_1}{\mu_1} \right)^2 \sum_{h=1}^{k_1-1} h(k_1 - h) \phi'_{1, h}{}^{(0)} - \left( 1 - \frac{\lambda_1}{\mu_1} \right)^2 \left( \frac{k_1}{k_2} \left( \frac{\lambda_1}{\mu_1} - k_1 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \right) \frac{\mu_2}{\mu_1} + \frac{\lambda_1}{\mu_1} \left( \frac{\mu_1}{\mu_2} + k_1 \right) \right) P'_0(\xi) \right\}. \tag{B.17}$$

PROOF: This equation follows from applying L'Hôpital's rule to  $\frac{A'^{(0)}(z, \xi)}{D(z)}$  and, hence, differentiating (B.10) and (B.11) with respect to  $z$ . Substitution of  $z = 1$  gives the desired result after, again, many algebraic manipulations. ■

Finally, we are ready to present the main result of this appendix, which is the proof of Eq. (B.1). Using Lemma B.1, and Eqs. (B.12) and (B.13), we can write the left-hand side of Eq. (B.1) as

$$\begin{aligned} & \mu_2 \frac{\lambda_1}{\mu_1} \left( k_2 X' - \frac{\lambda_1}{\mu_1} \left( \frac{k_2 - 1}{2} \right) \left( 1 - \frac{\lambda_1}{\mu_1} \right) P_0''(\xi) \right) \\ & - \mu_2 \left( 1 - \frac{\lambda_1}{\mu_1} \right) \left[ k_1 \frac{\mu_2}{\mu_1} X' + k_1 \tilde{Q}_1''^{(0)}(1, \xi) - \frac{\mu_2}{\mu_1} \left( 1 - \frac{\lambda_1}{\mu_1} \right) \sum_{h=1}^{k_1} (k_1 - h + 1) \tilde{Q}_h''^{(0)}(1, \xi) \right], \end{aligned}$$

where  $X'$  is the derivative of  $X$  with respect to  $\xi$ . Using Lemma B.3 and Lemma B.2 (and Eq. (A.6) to determine  $\tilde{Q}_h''^{(0)}(1, \xi)$  for  $h = 1, \dots, k_1$ ) we can show that the above expression reduces to  $\frac{\lambda_1 \mu_2^2}{\mu_1^2} P_0''(\xi)$ .