Waiting times in queueing networks with a single shared server*

M.A.A. Boon† marko@win.tue.nl
R.D. van der Mei ‡ mei@cwi.nl
E.M.M. Winands § e.m.m.winands@uva.nl

October 29, 2012

Abstract

We study a queueing network with a single shared server that serves the queues in a cyclic order. External customers arrive at the queues according to independent Poisson processes. After completing service, a customer either leaves the system or is routed to another queue. This model is very generic and finds many applications in computer systems, communication networks, manufacturing systems, and robotics. Special cases of the introduced network include well-known polling models, tandem queues, systems with a waiting room, multi-stage models with parallel queues, and many others. A complicating factor of this model is that the internally rerouted customers do not arrive at the various queues according to a Poisson process, causing standard techniques to find waiting-time distributions to fail. In this paper we develop a new method to obtain exact expressions for the Laplace-Stieltjes transforms of the steady-state waiting-time distributions. This method can be applied to a wide variety of models which lacked an analysis of the waiting-time distribution until now.

Keywords: queueing network, waiting times, customer routing, shared server, polling

Mathematics Subject Classification: 60K25, 90B22

1 Introduction

In this paper we study a queueing network served by a single shared server that visits the queues in a cyclic order. Customers from the outside arrive at the queues according to independent Poisson processes, and the service time and switch-over time distributions are general. After receiving service at queue i, a customer is either routed to queue j with

*The research was done in the framework of the BSIK/BRICKS project, the European Network of Excellence Euro-NF, and of the project "Service Optimization and Quality" (SeQual), funded by the Dutch agency SenterNovem.

†Eurandom and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands

‡Department of Mathematics, Section Stochastics, VU University, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands and Centre for Mathematics and Computer Science (CWI), 1098 SJ Amsterdam, The Netherlands

§University of Amsterdam, Korteweg-de Vries Institute for Mathematics, Science Park 904, 1098 XH Amsterdam, The Netherlands

probability $p_{i,j}$, or leaves the system with probability $p_{i,0}$. We consider systems with mixtures of gated and exhaustive service. This model can be seen as an extension of the standard polling model (in which customers always leave the system upon completion of their service) by customer routing. Yet another view is provided by the notion that the system is a Jackson network with a dedicated server for each queue with the additional complexity that only one server can be active in the network simultaneously.

The possibility of re-routing of customers further enhances the already-extensive modelling capabilities of polling models, which find applications in diverse areas such as computer systems, communication networks, logistics, flexible manufacturing systems, robotics systems, production systems and maintenance systems (see, for example, [5, 18, 22, 32] for overviews). Applications of the introduced type of customer routing can be found in many of these areas. In this regard, we would like to mention a manufacturing system where products undergo service in a number of stages or in the context of rework [17], a Ferry based Wireless Local Area Network (FWLAN) in which nodes can communicate with each other or with the outer world via a message ferry [20], a dynamic order picking system where the order picker drops off the picked items at the depot where sorting of the items is performed [16], and an internal mail delivery system where a clerk continuously makes rounds within the offices to pick up, sort and deliver mail [27].

In the past many papers have been published on special cases of the current network. In some of these papers distributional results are derived as well; the techniques used do, however, not allow for extension to the general setting of the current paper. Some special case configurations are standard polling systems [32], tandem queues [23, 34], multi-stage queueing models with parallel queues [19], feedback vacation queues [9, 33], symmetric feedback polling systems [31, 33], systems with a waiting room [1, 30], and many others. In conclusion, one can say that the present research can be seen as a unifying analysis of the waiting-time distribution for a wide variety of queueing models.

The main contribution of this paper is the derivation of waiting-time distributions in queueing networks with a single roving server via the development of a new method. For this model we derive the Laplace-Stieltjes transform of the waiting-time distribution of an arbitrary (internally rerouted, or external) customer. Due to this intrinsic complexity of the model, studies in the past were restricted to queue lengths and *mean* delay figures (see [6, 27, 28, 29]). A complicating, yet interesting, factor is that the combined process of internal and external arrivals violates the classical assumption of Poisson (or even renewal) arrivals, implying that traditional methods are not applicable. The basic idea behind the new method is that we explicitly compute a priori all *future* service requirements upon arrival of a newly arriving customer. In doing so the prerequisites of the distributional form of Little's Law are overcome.

An important feature of the newly developed technique is that it can be applied to a myriad of models which lacked an analysis of the waiting-time distribution until now. One could apply the framework (possibly after some minor modifications) to obtain distributional results in all of the aforementioned special cases of the studied system [1, 9, 19, 23, 30, 31, 32, 33, 34] but also, for example, in a closed network [2], in an $M/G/1$ queue with permanent and transient customers [8], in a network with permanent and transient customers [3], or in a polling model with arrival rates that depend on the location of the server [4, 7]. Although we study a continuous-time cyclic system with gated or exhaustive service in each queue, we may extend all results - without complicating the analysis - to discrete time, to periodic polling, to batch

arrivals, or to systems with different branching-type service disciplines such as globally gated service.

The structure of the present paper is as follows. In Section 2, we introduce the model and notation. Section 3 analyses the waiting-time distribution of an arbitrary customer for gated service. In Section 4 we study the system with mixtures of gated and exhaustive service. In the penultimate section, we present some examples which show the wide range of applicability of the studied model. The final section of this paper contains a brief discussion.

2 Model description and notation

We consider a queueing network consisting of $N \geq 1$ infinite buffer queues $Q_1, \ldots, Q_N$. External customers arrive at Q_i according to a Poisson arrival process with rate λ_i, and have a generally distributed service requirement B_i at Q_i, with mean value $b_i := \mathbb{E}[B_i]$. In general we denote the Laplace-Stieltjes Transform (LST) or Probability Generating Function (PGF) of a random variable X with $\widetilde{X}(\cdot)$. The queues are served by a single server in cyclic order. Whenever the server switches from Q_i to Q_{i+1}, a random switch-over time R_i is incurred, with mean r_i. The cycle time C_i is the time between successive moments when the server arrives at Q_i. The total switch-over time in a cycle is denoted by $R = \sum_{i=1}^{N} R_i$, and its first two moments are $r := \mathbb{E}[R]$ and $r^{(2)} := \mathbb{E}[R^2]$. Indices throughout the paper are modulo N, so Q_{1-N} and Q_{N+1} both refer to Q_1. All service times and switch-over times are mutually independent. This queueing network can be modelled as a *polling system* with the specific feature that it allows for routing of the customers: upon completion of service at Q_i, a customer is either routed to Q_j with probability $p_{i,j}$, or leaves the system with probability $p_{i,0}$. Note that $p_{i,0}$ should be greater than 0 for at least one queue, to make sure that customers can leave the system eventually. Moreover, note that $\sum_{j=0}^{N} p_{i,j} = 1$ for all i, and that the transition of a customer from Q_i to Q_j takes no time. Since we consider the gated and exhaustive service disciplines, the model under consideration has a branching structure, which is discussed in more detail by Foss [15] in the context of queueing models, and by Resing [26] more specifically in the context of polling systems. The total arrival rate at Q_i is denoted by γ_i, which is the unique solution of the following set of linear equations:

$$\gamma_i = \lambda_i + \sum_{j=1}^{N} \gamma_j p_{j,i}, \qquad i = 1, \ldots, N.$$

The offered load to Q_i is $\rho_i := \gamma_i b_i$ and the total load is $\rho := \sum_{i=1}^{N} \rho_i$. We assume that the system is stable, which means that ρ should be less than one (see [29]).

3 Gated service

In the present section we study the waiting-time distribution of an arbitrary customer for a system in which each queue receives gated service, which means that only those customers present at the server's arrival at Q_i will be served before the server switches to the next queue. We define the waiting time W_i of an arbitrary customer in Q_i as the time between his arrival at this queue and the moment at which his service starts. As far as waiting times are

concerned, a customer that is routed to another queue, say Q_j, upon his service completion is regarded as a new customer with waiting time W_j. The waiting-time distribution is found by conditioning on the numbers of customers present in each queue at an arrival epoch. To this end, we study the joint queue-length distribution at several embedded epochs in Section 3.1. In Sections 3.2 and 3.3 we use these results to successively derive the cycle-time distribution and the waiting-time distributions of internally rerouted customers and external customers.

3.1 The joint queue-length distributions

Sidi et al. [29] derive the PGFs of the joint queue-length distributions in all N queues at visit beginnings, visit completions, and at arbitrary points in time. In order to keep this manuscript self-contained, we briefly recapitulate their approach, as it forms the starting point of our novel method to find the waiting time LSTs. There is one important adaptation that we have to make, which will prove essential for finding waiting time LSTs. We consider not only the customers in all N queues, but we distinguish between customers standing *in front of* the gate and customers standing *behind* the gate (meaning that they will be served in the next cycle). Hence, we introduce the $N+1$ dimensional vector $\mathbf{z} = (z_1, \ldots, z_N, z_G)$. The element z_i, $i = 1, \ldots, N$, in this vector corresponds to customers in Q_i standing in front of the gate. The element z_G at position $N+1$ is only used during visit periods. During V_j, the visit period of Q_j, it corresponds to customers standing behind the gate in Q_j. This makes the analysis of systems with gated service slightly more involved than systems with exhaustive service (discussed in the next section). Before studying the joint queue-length distributions, we briefly introduce some convenient notation:

$$\begin{aligned}
\Sigma(\mathbf{z}) &= \sum_{j=1}^{N} \lambda_j (1 - z_j), \\
\Sigma_i(\mathbf{z}) &= \lambda_i (1 - z_G) + \sum_{j \neq i} \lambda_j (1 - z_j), \\
P_i(\mathbf{z}) &= p_{i,0} + p_{i,i} z_G + \sum_{j \neq i} p_{i,j} z_j.
\end{aligned}$$

Visit beginnings and completions. A cycle consists of N visit periods, V_i, each of which is followed by a switch-over time R_i, for $i = 1, \ldots, N$. A cycle C_i starts with a visit to Q_i and consists of the periods $V_i, R_i, V_{i+1}, \ldots, V_{i+N-1}, R_{i+N-1}$. Let P denote any of these periods. We denote the joint queue length PGF at the *beginning* of P as $\widetilde{LB}^{(P)}(\mathbf{z})$. The equivalent at the *completion* of period P is denoted by $\widetilde{LC}^{(P)}(\mathbf{z})$. Since the gated service discipline is a so-called *branching-type* service discipline (see [26]), we can express each of these functions in terms of $\widetilde{LB}^{(V_i)}(\mathbf{z})$, for any $i = 1, \ldots, N$. These relations, which are sometimes called *laws of motion*, are given below.

$$\widetilde{LC}^{(V_i)}(\mathbf{z}) = \widetilde{LB}^{(V_i)}\Big(z_1,\ldots,z_{i-1},\widetilde{B}_i\big(\Sigma_i(\mathbf{z})\big)P_i(\mathbf{z}),z_{i+1},\ldots,z_N,z_G\Big), \tag{3.1a}$$

$$\widetilde{LB}^{(R_i)}(\mathbf{z}) = \widetilde{LC}^{(V_i)}(z_1,\ldots,z_N,z_i), \tag{3.1b}$$

$$\widetilde{LC}^{(R_i)}(\mathbf{z}) = \widetilde{LB}^{(R_i)}(\mathbf{z})\widetilde{R}_i\Big(\Sigma(\mathbf{z})\Big), \tag{3.1c}$$

$$\widetilde{LB}^{(V_{i+1})}(\mathbf{z}) = \widetilde{LC}^{(R_i)}(\mathbf{z}), \tag{3.1d}$$

$$\vdots$$

$$\widetilde{LB}^{(V_{i+N})}(\mathbf{z}) = \widetilde{LC}^{(R_{i+N-1})}(\mathbf{z}). \tag{3.1e}$$

Note the subtle difference between $\widetilde{LC}^{(V_i)}(\mathbf{z})$ and $\widetilde{LB}^{(R_i)}(\mathbf{z})$, due to the fact that the gate in Q_i is removed after the completion of V_i, causing type G customers to become type i customers. In steady-state we have that $\widetilde{LB}^{(V_{i+N})}(\mathbf{z}) = \widetilde{LB}^{(V_i)}(\mathbf{z})$, implying that we have obtained a recursive relation for $\widetilde{LB}^{(V_i)}(\mathbf{z})$. Resing [26] shows how a clever definition of immigration and offspring generating functions can be used to find an explicit expression for $\widetilde{LB}^{(V_i)}(\mathbf{z})$. For reasons of compactness we refrain from doing so in the present paper. Instead we want to point out that the recursive relation obtained from (3.1a)-(3.1e) can be differentiated with respect to the variables $z_1,\ldots,z_N,z_G$. The resulting set of equations, which are called the *buffer occupancy equations* in the polling literature, can be used to compute the moments of the queue-length distributions at all visit beginnings and completions.

Service beginnings and completions. We denote the joint queue length PGF at *service* beginnings and completions in Q_j by respectively $\widetilde{LB}^{(B_j)}(\mathbf{z})$ and $\widetilde{LC}^{(B_j)}(\mathbf{z})$. Since a customer may be routed to another queue upon his service completion, we define $\widetilde{LC}^{(B_j)}(\mathbf{z})$ as the PGF of the joint queue-length distribution right *after* the tagged customer in Q_j has received service (implying that he is no longer present in Q_j), but *before* the moment that he may join another queue (even though these two epochs take place in a time span of length zero). Eisenberg [14] observed the following relation, albeit in a slightly different model:

$$\widetilde{LB}^{(V_i)}(\mathbf{z}) + \gamma_i\mathbb{E}[C]\widetilde{LC}^{(B_i)}(\mathbf{z})P_i(\mathbf{z}) = \widetilde{LC}^{(V_i)}(\mathbf{z}) + \gamma_i\mathbb{E}[C]\widetilde{LB}^{(B_i)}(\mathbf{z}). \tag{3.2}$$

Equation (3.2) is based on the observation that each visit beginning coincides with either a service beginning, or a visit completion (if no customer was present). Similarly, each visit completion coincides with either a visit beginning or a service completion. The long-run ratio between the number of visit beginnings/completions and service beginnings/completions in Q_i is $\gamma_i\mathbb{E}[C]$, with $\mathbb{E}[C] = \mathbb{E}[C_i] = r/(1-\rho)$. The distribution of the cycle time is given in the next subsection.

Furthermore, Eisenberg observes the following simple relation between the joint queue-length distribution at service beginnings and completions:

$$\widetilde{LC}^{(B_i)}(\mathbf{z}) = \widetilde{LB}^{(B_i)}(\mathbf{z})\widetilde{B}_i\big(\Sigma_i(\mathbf{z})\big)/z_i. \tag{3.3}$$

Substitution of (3.3) in (3.2) gives an equation which can be solved to express $\widetilde{LB}^{(B_i)}(\mathbf{z})$ in $\widetilde{LB}^{(V_i)}(\mathbf{z})$ and $\widetilde{LC}^{(V_i)}(\mathbf{z})$.

Arbitrary moments. The PGF of the joint queue-length distribution at arbitrary moments, denoted by $\widetilde{L}(\mathbf{z})$, is found by conditioning on the period in the cycle during which the system is observed $(V_1, R_1, \ldots, V_N, R_N)$.

$$\widetilde{L}(\mathbf{z}) = \frac{1}{\mathbb{E}[C]} \sum_{j=1}^{N} \Big(\mathbb{E}[V_j] \widetilde{L}^{(V_j)}(\mathbf{z}) + r_j \widetilde{L}^{(R_j)}(\mathbf{z}) \Big), \tag{3.4}$$

with $\mathbb{E}[V_j] = \rho_j \mathbb{E}[C]$. In (3.4) the functions $\widetilde{L}^{(V_j)}(\mathbf{z})$ and $\widetilde{L}^{(R_j)}(\mathbf{z})$ denote the PGFs of the joint queue-length distributions at an arbitrary moment during V_j and R_j respectively:

$$\widetilde{L}^{(V_j)}(\mathbf{z}) = \widetilde{LB}^{(B_j)}(\mathbf{z}) \frac{1 - \widetilde{B}_j\big(\Sigma_j(\mathbf{z})\big)}{b_j \Sigma_j(\mathbf{z})}, \tag{3.5}$$

$$\widetilde{L}^{(R_j)}(\mathbf{z}) = \widetilde{LB}^{(R_j)}(\mathbf{z}) \frac{1 - \widetilde{R}_j\big(\Sigma(\mathbf{z})\big)}{r_j \Sigma(\mathbf{z})}. \tag{3.6}$$

The interpretation of (3.5) and (3.6) is that the queue length vector at an arbitrary time point in V_j or R_j is the sum of those customers that were present at the beginning of that service/switch-over time, plus vector of the customers that have arrived during the elapsed part of the service/switch-over time. For more details about the joint queue length and workload distributions for general branching-type service disciplines (in the context of polling systems, but also applicable to our model) we refer to Boxma et al. [11].

3.2 Cycle-time distributions

In the remainder of this paper we present new results for the model introduced in Section 2. We start by analysing the distributions of the cycle times C_i, $i = 1, \ldots, N$. The idea behind the following analysis is to condition on the number of customers present in each queue at the beginning of C_i (and, hence, of V_i). The cycle will consist of the service of all of these customers, plus all switch-over times $R_i, \ldots, R_{i+N-1}$, plus the services of all customers that enter during these services and switch-over times *and* will be served *before* the next visit beginning to Q_i. The cycle time for polling systems without customer routing is discussed in Boxma et al. [10]. However, as it turns out, the analysis is severely complicated by the fact that customers may be routed to another queue and be served again (even multiple times) during the same cycle.

From branching theory we adopt the term *descendants* of a certain (tagged) customer to denote all customers that arrive (in all queues) during the service of this tagged customer, plus the customers arriving during their service times, and so on. If, upon his service completion, a customer is routed to another queue, we also consider him as his own descendant. We define $B^*_{k,i}$, $i = 1, \ldots, N; k = 0, \ldots, N$, as the service time of a type $i-k$ (which is understood as $N+i-k$ if $i \leq k$) customer at Q_{i-k}, plus the service times of all of his descendants that will be served before or during the next visit of the server to Q_i. The special case $B^*_{0,i}$ is simply the service time of a type i customer, $i = 1, \ldots, N$. A formal definition in terms of LSTs is given below:

$$\widetilde{B}^*_{k,i}(\omega) = \widetilde{B}_{i-k}\Big(\omega + \sum_{j=0}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j,i}(\omega)\big)\Big) \widetilde{P}^*_{k,i}(\omega), \quad k = 0, 1, \ldots, N; i = 1, \ldots, N, \tag{3.7}$$

where

$$\widetilde{P}^*_{k,i}(\omega) = 1 - \sum_{j=0}^{k-1} p_{i-k,i-j}\big(1 - \widetilde{B}^*_{j,i}(\omega)\big), \qquad k = 0, 1, \ldots, N; i = 1, \ldots, N. \quad (3.8)$$

For a type $i-k$ customer, $P^*_{k,i}$ accounts for the service times of his descendants that are caused by the fact that he may be routed to another queue upon his service completion.

A similar function should be defined for the switch-over times:

$$\widetilde{R}^*_{k,i}(\omega) = \widetilde{R}_{i-k}\Big(\omega + \sum_{j=0}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j,i}(\omega)\big)\Big), \qquad k = 0, 1, \ldots, N; i = 1, \ldots, N. \quad (3.9)$$

Note that, compared to (3.7), no term $\widetilde{P}^*_{k,i}(\omega)$ is required because no routing takes place at the end of a switch-over time.

Finally, we define the following $N+1$ dimensional vectors:

$$\mathbf{B_{k,i}} = \big(1, \ldots, 1, \widetilde{B}^*_{k,i}(\omega), 1, \ldots, 1\big), \qquad k = 0, 1, \ldots, N-1; i = 1, \ldots, N, \quad (3.10)$$

$$\mathbf{B_{N,i}} = \big(1, \ldots, 1, \widetilde{B}^*_{0,i}(\omega)\big), \qquad i = 1, \ldots, N, \quad (3.11)$$

with $\widetilde{B}^*_{k,i}(\omega)$ at position $i-k$ in (3.10) (or position $N+i-k$ if $k \geq i$), and $\widetilde{B}^*_{0,i}(\omega)$ at position $N+1$ in (3.11). We use $\bigotimes$ to denote the element-wise multiplication of vectors.

Proposition 3.1 The LST of the distribution of the cycle time C_i is given by

$$\widetilde{C}_i(\omega) = \widetilde{LB}^{(V_i)}\big(\bigotimes_{k=0}^{N-1} \mathbf{B_{k,i-1}}\big) \prod_{k=0}^{N-1} \widetilde{R}^*_{k,i-1}(\omega), \qquad i = 1, \ldots, N. \quad (3.12)$$

The interpretation of (3.12) is that the length of a cycle starting with a visit to Q_i is the sum of the *extended* service times of all customers present at the beginning of the cycle, and the sum of all *extended* switch-over times during the cycle. By extended service time (switch-over time) we refer to a service time (switch-over time) plus the service times of all customers that arrive during this service time (switch-over time) in one of the queues that are yet to be served during the remainder of the cycle, and all of their descendants that will be served before the end of the cycle.

Proof: To prove Proposition 3.1 we keep track of all the customers that will be served during one cycle. We condition on the numbers of customers present in each queue at the beginning of C_i, denoted by $n_1, \ldots, n_N$. Note that there are no gated customers present at this moment, because the gate has been removed at the beginning of the last switch-over time of the previous cycle. A cycle C_i consists of:

1. the service of all customers present at the beginning of the cycle,
2. all of their descendants that will be served before the start of the next cycle (i.e., before the next visit to Q_i),
3. the switch-over times $R_1, \ldots, R_N$,

4. all customers arriving during these switch-over times that will be served before the start of the next cycle,

5. all of their descendants that will be served before the start of the next cycle.

We define S_j for $j = 1, \ldots, N$, as the service time of a type j customer plus the service times of all of his descendants that will be served during (the remaining part of) C_i. Since the service discipline is gated at all queues, we have:

$$S_j = B_j + \sum_{k=j+1}^{i-1} \sum_{l=1}^{N_k(B_j)} S_{k_l} + \begin{cases} S_m & \text{for } m = j+1, \ldots, i-1, \text{ w.p. } p_{j,m}, \\ 0 & \text{w.p. } 1 - \sum_{m=j+1}^{i-1} p_{j,m}, \end{cases} \tag{3.13}$$

where $N_k(T)$ denotes the number of arrivals in Q_k during a (possibly random) period of time T, and S_{k_l} is a sequence of (independent) extended service times S_k. Note that S_j depends on i, although we have chosen to hide this for presentational purposes. The gated service discipline is reflected in the fact that only customers arriving in (or rerouted to) $Q_{j+1}, \ldots, Q_{i-1}$ are being served during the residual part of C_i. It can easily be shown that the LST of S_{i-k} is $\widetilde{B}^*_{k-1,i-1}(\omega)$ for $k = 1, \ldots, N$. Note that the first summation in (3.13) is cyclic, which may sometimes cause confusion (for example if $j = i-1$, when this is supposed to be a summation over zero terms). Avoiding this (possible) confusion is the main reason that we have chosen to define $\widetilde{B}^*_{k,i}(\omega)$, $\widetilde{P}^*_{k,i}(\omega)$ and $\widetilde{R}^*_{k,i}(\omega)$ relative to queue i (k steps backward in time).

Using this branching way of looking at the cycle time, we can express C_i in terms of $R_1, \ldots, R_N$ and $S_1, \ldots, S_N$. First, however, we derive the following intermediate result.

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\omega R_{i-k}} \prod_{j=i-k+1}^{i-1} \prod_{l=1}^{N_j(R_j)} \mathrm{e}^{-\omega S_{j_l}} \right] &= \widetilde{R}_{i-k}\big(\omega + \sum_{j=i-k+1}^{i-1} \lambda_j (1 - \mathbb{E}[\mathrm{e}^{-\omega S_j}])\big) \\ &= \widetilde{R}^*_{k-1,i-1}(\omega). \end{aligned}$$

Now, introducing the shorthand notation $n_1, \ldots, n_N$ for the event that the numbers of customers at the beginning of C_i in queues $1, \ldots, N$ are respectively $n_1, \ldots, n_N$, we can find the cycle time LST conditional on this event.

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\omega C_i} \,|\, n_1, \ldots, n_N\right] &= \mathbb{E}\left[\exp\Big(-\omega \sum_{j=i-N}^{i-1} \big(\sum_{l=1}^{n_j} S_{j_l} + R_j + \sum_{k=j+1}^{i-1} \sum_{l=1}^{N_k(R_j)} S_{k_l} \big) \Big) \right] \\ &= \mathbb{E}\left[\prod_{j=i-N}^{i-1} \left(\prod_{l=1}^{n_j} \mathrm{e}^{-\omega S_{j_l}} \right) \mathrm{e}^{-\omega R_j} \prod_{k=j+1}^{i-1} \prod_{l=1}^{N_k(R_j)} \mathrm{e}^{-\omega S_{k_l}} \right] \\ &= \prod_{j=i-N}^{i-1} \left(\prod_{l=1}^{n_j} \mathbb{E}\left[\mathrm{e}^{-\omega S_{j_l}}\right] \right) \prod_{j=i-N}^{i-1} \mathbb{E}\left[\mathrm{e}^{-\omega R_j} \prod_{k=j+1}^{i-1} \prod_{l=1}^{N_k(R_j)} \left(\mathrm{e}^{-\omega S_{k_l}}\right) \right] \\ &= \left(\prod_{k=1}^{N} \widetilde{B}^*_{k-1,i-1}(\omega)^{n_{i-k}} \right) \prod_{k=1}^{N} \widetilde{R}^*_{k-1,i-1}(\omega). \end{aligned}$$

Equation (3.12) follows after deconditioning. □

Remark 3.2 Because of our main interest in the waiting-time distributions, we have followed quite an elaborate path to find the LST of the cycle-time distribution. However, if one is merely interested in a quick way to find $\widetilde{C}_i(\omega)$, a more efficient approach can be used. One of the most efficient ways to find $\widetilde{C}_i(\omega)$ is to distinguish between customers that arrive from outside the network (external customers) and internally rerouted customers (internal customers). One can straightforwardly adapt the laws of motion (3.1a)-(3.1e) to find an expression for $\widetilde{LB}^{(V_i)'}(z_1^E, z_1^I, \ldots, z_N^E, z_N^I)$. Just like $\widetilde{LB}^{(V_i)}(z_1, \ldots, z_N, z_G)$, $\widetilde{LB}^{(V_i)'}(z_1^E, z_1^I, \ldots, z_N^E, z_N^I)$ stands for the PGF of the joint queue length at the beginning of V_i, but now we distinguish between external and internal customers in each queue (indicated by z_j^E and z_j^I). Since external customers arrive in Q_i according to a Poisson process with intensity λ_i, one can apply the distributional form of Little's Law (see, for example, Keilson and Servi [21]) to the *external* customers in Q_i:

$$\widetilde{C}_i(\omega) = \widetilde{LB}^{(V_i)'}(1, \ldots, 1, 1 - \omega/\lambda_i, 1, \ldots, 1), \qquad i = 1, \ldots, N.$$

3.3 Waiting-time distributions

In this subsection we find the LSTs of W_i^E and W_i^I, the waiting-time distributions of arbitrary external and internal customers in Q_i, and use them to obtain the LST of W_i, the waiting time of an arbitrary customer. Recall that the waiting time W_i of an arbitrary customer in Q_i is the time between his arrival at this queue and the moment at which his service starts. Hence, even if a customer is routed to the same queue multiple times, each visit to this queue invokes a new waiting time. We stress that common methods used in the polling literature to find waiting time LSTs cannot be applied in our queueing network, because they rely heavily on the assumption that *every* customer in the system has arrived according to a Poisson process. Since this assumption is violated in our model, we have developed a novel approach to find the waiting time LST of an arbitrary customer in our network. The joint queue-length distributions at various epochs, as discussed in Subsection 3.1, play an essential role in the analysis. First we focus on the waiting times of internal customers, then we discuss the waiting times of external customers.

Internal customers. The arrival epoch of an internal customer always coincides with a service completion. Hence, we condition on the joint queue length and the arrival epoch of an internal customer to find his waiting time LST. The waiting time of an internal customer *given that* he arrives in Q_i after a service completion at Q_{i-k} is denoted by $WC_i^{(B_{i-k})}$ $(i, k = 1, \ldots, N)$. To find $WC_i^{(B_{i-k})}$, we only have to compute the probability that an arbitrary internal customer in Q_i arrives after a service completion at Q_{i-k}. The mean number of customers (internal plus external) present at the beginning of V_{i-k} at Q_{i-k} is $\gamma_{i-k}\mathbb{E}[C]$. Each of these customers joins Q_i upon his service completion with probability $p_{i-k,i}$. This observation combined with the fact that the mean number of *internal* customers arriving at Q_i during the course of one cycle is $(\gamma_i - \lambda_i)\mathbb{E}[C]$, leads to the following result:

$$\widetilde{W}_i^I(\omega) = \sum_{k=1}^{N} \frac{\gamma_{i-k} p_{i-k,i}}{\gamma_i - \lambda_i} \widetilde{WC}_i^{(B_{i-k})}(\omega), \qquad i = 1, \ldots, N. \tag{3.14}$$

As a consequence, the problem of finding $\widetilde{W}_i^I(\cdot)$ is reduced to finding $\widetilde{WC}_i^{(B_{i-k})}(\omega)$ for all $i,k=1,\ldots,N$.

For notational reasons we first introduce the following $N+1$ dimensional vectors, which will appear several times in this section:

$$\mathbf{B}_{\boldsymbol{k,i}}^{\mathbf{G}} = \begin{cases} \mathbf{B}_{\boldsymbol{0,i}} & \text{if } k<0, \\ \mathbf{B}_{\boldsymbol{0,i}} \bigotimes_{j=0}^{k-1} \mathbf{B}_{\boldsymbol{j,i-1}} & \text{if } k=1,\ldots,N, \\ \mathbf{B}_{\boldsymbol{N,i}} \bigotimes_{j=0}^{N-1} \mathbf{B}_{\boldsymbol{j,i-1}} & \text{if } k=N, \end{cases}$$

for $i=1,\ldots,N$. Again, we use $\bigotimes$ to denote the element-wise multiplication of vectors.

Proposition 3.3 We have

$$\widetilde{WC}_i^{(B_{i-k})}(\omega) = \widetilde{LC}^{(B_{i-k})}\big(\mathbf{B}_{\boldsymbol{k,i}}^{\mathbf{G}}\big) \prod_{j=0}^{k-1} \widetilde{R}_{j,i-1}^*(\omega), \tag{3.15}$$

for $i,k=1,\ldots,N$.

Proof: The key observation in the proof of Proposition 3.3 is that an arrival of an internally rerouted customer always coincides with some service completion. For this reason, we consider the system right after the service completion at, say, Q_j ($j=1,\ldots,N$). We compute the waiting time LST of a customer routed to Q_i after being served in Q_j, conditional on the numbers of customers of each type (now *including* gated customers) present at the arrival epoch (*not* including the arriving customer himself). We denote by $n_1,\ldots,n_N,n_G$ the event that the numbers of customers of all types are respectively $n_1,\ldots,n_N,n_G$. Let $n_{iG} := n_i$ if $i \neq j$, and $n_{iG} := n_G$ if $i=j$. Note that the type G customers are located behind the gate in Q_j, and that the customer routed to Q_i only has to wait for these customers in case $i=j$. The waiting time of the tagged customer consists of:

1. the service of all n_j customers in front of the gate in Q_j at the arrival epoch,
2. the service of all $n_{j+1},\ldots,n_{i-1}$ customers present in $Q_{j+1},\ldots,Q_{i-1}$ at the arrival epoch,
3. all of the descendants of the previously mentioned customers that will be served before the next visit to Q_i,
4. if $i \neq j$, the service of all n_{iG} customers present in Q_i at the arrival epoch; if $i=j$, the service of all n_{iG} gated customers present in Q_i at the arrival epoch,
5. the switch-over times $R_j,\ldots,R_{i-1}$,
6. all customers arriving during these switch-over times that will be served before the next visit to Q_i,
7. all of their descendants that will be served before the next visit to Q_i.

We denote the waiting time of an internal customer conditional on the event that he arrives in Q_i after being served in Q_j, *and conditional on the event that the numbers of customers of all types at the arrival epoch are respectively* $n_1, \ldots, n_N, n_G$, by $WC_i^{(B_j)'}$. Just like in the proof of Proposition 3.1, we can express $WC_i^{(B_j)'}$ in terms of $R_1, \ldots, R_N$ and $S_1, \ldots, S_N$:

$$WC_i^{(B_j)'} = \sum_{k=j}^{i-1} \left[\sum_{l=1}^{n_k} S_{k_l} + R_k + \sum_{l=k+1}^{i-1} \sum_{m=1}^{N_l(R_k)} S_{l_m} \right] + \sum_{l=1}^{n_{iG}} B_{i,l}. \tag{3.16}$$

Taking the LST of (3.16) leads to (3.15) after deconditioning. The derivation proceeds along the exact same lines as in the proof of Proposition 3.1, and is therefore omitted.

□

External customers. External customers arrive in Q_i according to a Poisson process with intensity λ_i. We distinguish between customers arriving during a switch-over time and customers arriving during a visit time. The waiting time of an external customer in Q_i *given that* he arrives during R_{i-k} is denoted by $W_i^{(R_{i-k})}$ $(i, k = 1, \ldots, N)$. Similarly, we use $W_i^{(V_{i-k})}$ to denote an external customer arriving in Q_i during V_{i-k}. The waiting time LST of an arbitrary external customer can be expressed in terms of $\widetilde{W}_i^{(R_{i-k})}(\cdot)$ and $\widetilde{W}_i^{(V_{i-k})}(\cdot)$:

$$\widetilde{W}_i^E(\omega) = \frac{1}{\mathbb{E}[C]} \sum_{k=1}^{N} \left(\mathbb{E}[V_{i-k}] \widetilde{W}_i^{(V_{i-k})}(\omega) + r_{i-k} \widetilde{W}_i^{(R_{i-k})}(\omega) \right), \qquad i = 1, \ldots, N. \tag{3.17}$$

We first focus on the waiting time of customers arriving during a switch-over time. Consider a tagged customer arriving in Q_i during R_{i-k}, $i, k = 1, \ldots, N$. Since the remaining part of the switch-over time is part of the waiting time of the arriving customer, it will turn out that we need the *joint* distribution of all customers present at the arrival epoch *and* the residual part of R_{i-k}, denoted by R_{i-k}^R. The PGF of the joint queue-length distribution at the arrival epoch is given by (3.6). Equation (3.6) is based on the observation that the number of customers in each queue at an arbitrary moment during R_{i-k} is simply the sum of the number of customers present at the beginning of R_{i-k} and the number of customers that have arrived during the elapsed (past) part of R_{i-k}, denoted by R_{i-k}^P. These random variables are independent. Hence, it is straightforward to adapt (3.6) to find the joint distribution of the queue lengths *and* residual part of R_{i-k}, using the following result from elementary renewal theory:

$$\widetilde{R}_j^{PR}(\omega_P, \omega_R) = \frac{\widetilde{R}_j(\omega_P) - \widetilde{R}_j(\omega_R)}{(\omega_R - \omega_P) r_j}, \qquad j = 1, \ldots, N,$$

with $\widetilde{R}_j^{PR}(\omega_P, \omega_R)$ denoting the LST of the joint distribution of past and residual switch-over time R_j. Hence,

$$\widetilde{L}^{(R_j)}(\mathbf{z}, \omega) = \widetilde{LB}^{(R_j)}(\mathbf{z}) \widetilde{R}_j^{PR}(\Sigma(\mathbf{z}), \omega), \tag{3.18}$$

where $\widetilde{L}^{(R_j)}(\mathbf{z}, \omega)$ denotes the PGF-LST of the joint distribution of the number of customers of each type at an arbitrary moment during R_j and the residual part of R_j. Obviously, there are no gated customers present during a switch-over time.

Consequently, and also using PASTA, we can find the waiting-time distribution by conditioning on the number of customers present at an arbitrary moment during R_{i-k} and on the residual switch-over time.

Proposition 3.4 We have

$$\widetilde{W}_i^{(R_{i-k})}(\omega) = \widetilde{R}_{i-k}^{PR}\Big(\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)+\lambda_i\big(1-\widetilde{B}_i(\omega)\big),\omega+\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)\Big)$$
$$\times\, \widetilde{LB}^{(R_{i-k})}\big(\mathbf{B}^{\mathbf{G}}_{\boldsymbol{k-1,i}}\big)\prod_{j=0}^{k-2}\widetilde{R}^*_{j,i-1}(\omega), \qquad i,k=1,\ldots,N, \tag{3.19}$$

Proof: We consider an arbitrary customer arriving in Q_i during R_j. Similar to the proofs of the preceding propositions in this section, we condition on the number of customers present in all queues at the arrival epoch, denoted by $n_1,\ldots,n_N$. As mentioned before, no gated customers are present during a switch-over time. However, we also condition on the residual length of R_j, denoted by t_R. The waiting time of the tagged customer consists of:

1. the service of all $n_{j+1},\ldots,n_{i-1}$ customers present at the arrival epoch in $Q_{j+1},\ldots,Q_{i-1}$,
2. the service of all their descendants that will be served before the start of the next visit to Q_i,
3. the service of all n_i customers present at the arrival epoch in Q_i,
4. the residual switch-over time t_R,
5. the switch-over times $R_{j+1},\ldots,R_{i-1}$,
6. the service of all customers arriving during $t_R, R_{j+1},\ldots,R_{i-1}$ that will be served before the start of the next visit to Q_i,
7. the service of all descendants of these customers that will be served before the start of the next visit to Q_i.

If we denote the waiting time of a type i customer arriving during R_j, *conditional on* $n_1,\ldots,n_N$ and t_R, by $W_i^{(R_j)'}$, we can summarise these items in the following formula:

$$W_i^{(R_j)'} = \sum_{k=j+1}^{i-1}\left[\sum_{l=1}^{n_k}S_{k_l}+R_k+\sum_{l=k+1}^{i-1}\sum_{m=1}^{N_l(R_k)}S_{l_m}\right]+\sum_{l=1}^{n_i}B_{i_l}+t_R+\sum_{l=j+1}^{i-1}\sum_{m=1}^{N_l(t_R)}S_{l_m}. \tag{3.20}$$

Taking the LST of (3.20) and using (3.18) leads to (3.19) after deconditioning. The derivation is not completely straightforward, but rather than providing it here, we refer to the proof of Proposition 3.5, which contains a similar derivation of a more complicated equation. □

Now we only need to determine $\widetilde{W}_i^{(V_{i-k})}(\cdot)$. Focussing on a tagged customer arriving in Q_i during the service of a customer in Q_{i-k}, for $i,k=1,\ldots,N$, we can find $\widetilde{W}_i^{(V_{i-k})}(\cdot)$ by conditioning on the number of customers in each queue at the arrival epoch and the residual service time. Similar to $\widetilde{R}_j^{PR}(\cdot)$, we define the LST of the joint distribution of past and residual service time B_j as

$$\widetilde{B}_j^{PR}(\omega_P,\omega_R) = \frac{\widetilde{B}_j(\omega_P)-\widetilde{B}_j(\omega_R)}{(\omega_R-\omega_P)b_j}, \qquad j=1,\ldots,N. \tag{3.21}$$

We can now use Equations (3.5) and (3.21) to find the PGF-LST of the joint distribution of the number of customers of each type present at an arbitrary moment during V_j and the residual service time of the customer that is being served at that moment:

$$\widetilde{L}^{(V_j)}(\mathbf{z},\omega) = \widetilde{LB}^{(B_j)}(\mathbf{z})\widetilde{B}_j^{PR}(\Sigma_j(\mathbf{z}),\omega). \tag{3.22}$$

Note that the customers arriving in Q_j during the elapsed part of B_j are gated customers.

Proposition 3.5 We have

$$\widetilde{W}_i^{(V_{i-k})}(\omega) = \widetilde{B}_{i-k}^{PR}\Big(\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)+\lambda_i\big(1-\widetilde{B}_i(\omega)\big),\omega+\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)\Big)$$
$$\times\,\widetilde{LB}^{(B_{i-k})}\big(\mathbf{B}^{\mathbf{G}}_{\boldsymbol{k,i}}\big)\prod_{j=0}^{k-1}\widetilde{R}^*_{j,i-1}(\omega)\times\frac{\widetilde{P}^*_{k-1,i-1}(\omega)}{\widetilde{B}^*_{k-1,i-1}(\omega)}, \tag{3.23}$$

for $i,k=1,\ldots,N$.

Proof: We denote by $n_1,\ldots,n_N,n_G$ the numbers of customers of all types present at the arrival epoch of the tagged customer. The residual part of the service time of the customer being served at this arrival epoch is denoted by t_R. Let $n_{iG}:=n_i$ if $i\neq j$, and $n_{iG}:=n_G$ if $i=j$. The waiting time of a type i customer arriving during V_j, conditional on $n_1,\ldots,n_N,n_G$ and the residual service time consists of the following components:

1. the service of n_j-1 customers in front of the gate in Q_j (We exclude the customer being served at the arrival epoch),
2. the service of all $n_{j+1},\ldots,n_{i-1}$ customers present in $Q_{j+1},\ldots,Q_{i-1}$,
3. all of the descendants of the previously mentioned customers that will be served before the next visit to Q_i,
4. if $i\neq j$, the service of all n_{iG} customers present in Q_i at the arrival epoch; if $i=j$, the service of all n_{iG} gated customers present in Q_i,
5. the switch-over times $R_j,\ldots,R_{i-1}$,
6. the residual service time t_R,
7. all customers arriving during t_R and $R_j,\ldots,R_{i-1}$ that will be served before the next visit to Q_i,
8. all of their descendants that will be served before the next visit to Q_i,
9. the (possible) future service of the customer being served at the arrival epoch, due to the fact that he may be routed to another queue that will be served before the next visit to Q_i,
10. the service of all descendants of this rerouted customer (Note that if he will be rerouted and served again, he will count as his own descendant).

More formally:

$$W_i^{(V_j)'} = \sum_{l=1}^{n_j-1} S_{j,l} + \sum_{k=j+1}^{i-1} \sum_{l=1}^{n_k} S_{k_l} + \sum_{l=1}^{n_{iG}} B_{i_l} + \sum_{k=j}^{i-1} \left[R_k + \sum_{l=k+1}^{i-1} \sum_{m=1}^{N_l(R_k)} S_{l_m} \right] + t_R + \sum_{l=j+1}^{i-1} \sum_{m=1}^{N_l(t_R)} S_{l_m} + \begin{cases} S_l & \text{for } l = j+1, \ldots, i-1, \text{ w.p. } p_{j,l}, \\ 0 & \text{w.p. } 1 - \sum_{l=j+1}^{i-1} p_{j,l}, \end{cases} . \tag{3.24}$$

We now show that Equation (3.23) follows from taking the LST:

$$\begin{aligned}
&\mathbb{E}[\mathrm{e}^{-\omega W_i^{(V_j)}} | n_1, \ldots, n_N, n_{iG}] \\
&= \mathbb{E}\left[\prod_{l=1}^{n_j-1} \mathrm{e}^{-\omega S_{j_l}} \prod_{m=j+1}^{i-1} \prod_{l=1}^{n_m} \mathrm{e}^{-\omega S_{m_l}} \right] \mathbb{E}\left[\prod_{l=1}^{n_{iG}} \mathrm{e}^{-\omega B_{i_l}} \right] \mathbb{E}\left[\prod_{m=j}^{i-1} \mathrm{e}^{-\omega \left(R_m + \sum_{l=m+1}^{i-1} \sum_{q=1}^{N_l(R_m)} S_{l_q} \right)} \right] \\
&\times \mathrm{e}^{-\omega t_R} \mathbb{E}\left[\prod_{l=j+1}^{i-1} \prod_{m=1}^{N_l(t_R)} \mathrm{e}^{-\omega S_{l_m}} \right] \left(\sum_{l=j+1}^{i-1} p_{j,l} \mathbb{E}\left[\mathrm{e}^{-\omega S_l} \right] + 1 - \sum_{l=j+1}^{i-1} p_{j,l} \right) \\
&= \mathbb{E}\left[\mathrm{e}^{-\omega S_j} \right]^{n_j-1} \prod_{m=j+1}^{i-1} \mathbb{E}\left[\mathrm{e}^{-\omega S_m} \right]^{n_m} \mathbb{E}\left[\mathrm{e}^{-\omega B_i} \right]^{n_{iG}} \prod_{m=j}^{i-1} \widetilde{R}_m\Big(\omega + \sum_{l=m+1}^{i-1} (1 - \mathbb{E}[\mathrm{e}^{-\omega S_l}]) \Big) \\
&\times \mathrm{e}^{-\omega t_R} \prod_{l=j+1}^{i-1} \sum_{m=0}^{\infty} \mathbb{E}[\mathrm{e}^{-\omega S_l}]^m \mathbb{P}[N_l(t_R) = m] \left(1 - \sum_{l=j+1}^{i-1} p_{j,l} \Big(1 - \mathbb{E}\left[\mathrm{e}^{-\omega S_l} \right] \Big) \right) \\
&= \widetilde{B}^*_{k-1,i-1}(\omega)^{n_{i-k}-1} \prod_{l=1}^{k-1} \widetilde{B}^*_{l-1,i-1}(\omega)^{n_{i-l}} \widetilde{B}_i(\omega)^{n_{iG}} \prod_{l=1}^{k} \widetilde{R}^*_{l-1,i-1}(\omega) \\
&\times \exp\left[-\Big(\omega + \sum_{l=j+1}^{i-1} (1 - \mathbb{E}[\mathrm{e}^{-\omega S_l}]) \Big) t_R \right] \widetilde{P}^*_{k-1,i-1}(\omega) \\
&= \widetilde{B}^*_{k-1,i-1}(\omega)^{n_{i-k}} \prod_{l=1}^{k-1} \widetilde{B}^*_{l-1,i-1}(\omega)^{n_{i-l}} \widetilde{B}_i(\omega)^{n_{iG}} \prod_{l=1}^{k} \widetilde{R}^*_{l-1,i-1}(\omega) \\
&\times \exp\left[-\Big(\omega + \sum_{l=1}^{k-1} (1 - \widetilde{B}^*_{l-1,i-1}(\omega)) \Big) t_R \right] \frac{\mathbb{P}_{k-1,i-1}(\omega)}{\widetilde{B}^*_{k-1,i-1}(\omega)},
\end{aligned}$$

where $k = i - j$ (or $k = N + i - j$ if $j \geq i$). Deconditioning of this expression leads to (3.23). □

Arbitrary customers. Finally, we present the main result of this section: the LST of the waiting-time distribution of an arbitrary customer in Q_i.

Theorem 3.6 The LST of the waiting-time distribution of an arbitrary customer in Q_i, if this queue receives gated service, is given by:

$$\widetilde{W}_i(\omega) = \frac{\gamma_i - \lambda_i}{\gamma_i} \widetilde{W}_i^I(\omega) + \frac{\lambda_i}{\gamma_i} \widetilde{W}_i^E(\omega), \qquad i = 1, \ldots, N, \tag{3.25}$$

where $\widetilde{W}_i^I(\omega)$ and $\widetilde{W}_i^E(\omega)$ are given by (3.14) and (3.17), respectively.

Proof: The result follows immediately after conditioning on the event that an arbitrary customer is an internal or external customer. □

4 Exhaustive service

In this section we study systems with mixtures of gated and exhaustive service, that is, some queues are served exhaustively whereas other queues receive gated service. We restrict ourselves to presenting the results, but for reasons of compactness we omit all proofs as they can be produced similar to the proofs in the previous section.

Throughout we use the index $e \in \{1, \ldots, N\}$ to refer to an arbitrary queue with exhaustive service, which means that customers are being served until the queue is empty. This means that, in contrast to gated service, customers arriving in Q_e *during* V_e will be served during that same visit period. This is true, even if the customer has just received service in Q_e and was routed back to Q_e again. To deal with this issue, we define an extended service time B_e^{exh} which is the total amount of service that a customer receives during a visit period V_e before being routed to another queue (or leaving the system), cf. [29]. As stated in [29], B_e^{exh} is the geometric sum, with parameter $p_{e,e}$, of independent random variables with the same distribution as B_e. The LST of B_e^{exh} is given by

$$\widetilde{B}_e^{\text{exh}}(\omega) = \frac{(1-p_{e,e})\widetilde{B}_e(\omega)}{1-p_{e,e}\widetilde{B}_e(\omega)}.$$

We denote a busy period of type e customers by BP_e. The PGF-LST of the joint distribution of a busy period and the number of customers served during this busy period satisfies the following equation:

$$\widetilde{BP}_e(z,\omega) = z\widetilde{B}_e^{\text{exh}}\big(\omega + \lambda_e(1-\widetilde{BP}_e(z,\omega))\big).$$

4.1 The joint queue-length distributions

Visit beginnings and completions. The laws of motion (3.1a)-(3.1e) have to be adapted if a queue receives exhaustive service. First we need to redefine $\Sigma_i(\mathbf{z})$ and $P_i(\mathbf{z})$ if Q_i is served exhaustively, and introduce $P_i^{\text{exh}}(\mathbf{z})$:

$$\begin{aligned}
\Sigma_e(\mathbf{z}) &= \sum_{j\neq e} \lambda_j(1-z_j),\\
P_e(\mathbf{z}) &= p_{e,0} + \sum_{j=1}^{N} p_{e,j}z_j,\\
P_e^{\text{exh}}(\mathbf{z}) &= \frac{p_{e,0}}{1-p_{e,e}} + \sum_{j\neq e}\frac{p_{e,j}}{1-p_{e,e}}z_j,
\end{aligned}$$

for all $e \in \{1, \ldots, N\}$ corresponding to queues with exhaustive service. The laws of motion now change accordingly:

$$\widetilde{LC}^{(V_e)}(\mathbf{z}) = \widetilde{LB}^{(V_e)}\Big(z_1, \ldots, z_{e-1}, \widetilde{BP}_e\big(P_e^{\text{exh}}(\mathbf{z}), \Sigma_e(\mathbf{z})\big), z_{e+1}, \ldots, z_N, 1\Big),$$
$$\widetilde{LB}^{(R_e)}(\mathbf{z}) = \widetilde{LC}^{(V_e)}(\mathbf{z}),$$

for any exhaustively served Q_e.

Service beginnings and completions. Eisenberg's relation (3.2) remains valid for queues with exhaustive service. Note that $P_e(\mathbf{z})$ should *not* be replaced by $P_e^{\text{exh}}(\mathbf{z})$ for exhaustive queues in (3.2)! Relation (3.3) should be slightly changed for queues with exhaustive service, since customers are not placed behind a gate:

$$\widetilde{LC}^{(B_e)}(\mathbf{z}) = \widetilde{LB}^{(B_e)}(\mathbf{z})\widetilde{B}_e\big(\Sigma(\mathbf{z})\big)/z_e.$$

Arbitrary moments. Equation (3.4) for the PGF of the joint queue-length distribution at arbitrary moments remains valid if some of the queues have exhaustive service. However, $\widetilde{L}^{(V_j)}(\mathbf{z})$ should be adapted for queues with exhaustive service by replacing gated customers with "ordinary" type e customers:

$$\widetilde{L}^{(V_e)}(\mathbf{z}) = \widetilde{LB}^{(B_e)}(\mathbf{z})\frac{1 - \widetilde{B}_e\big(\Sigma(\mathbf{z})\big)}{b_e\Sigma(\mathbf{z})}.$$

4.2 Cycle-time distributions

The fact that customers arriving in an exhaustively served queue, say Q_{i-k}, during V_{i-k} are served before the end of this visit period, requires changes in the definition of $\widetilde{B}^*_{k,i}(\omega)$.

$$\widetilde{B}^*_{k,i}(\omega) = \widetilde{BP}_{i-k}\Big(\widetilde{P}^*_{k,i}(\omega), \omega + \sum_{j=0}^{k-1}\lambda_{i-j}(1 - \widetilde{B}^*_{j,i}(\omega))\Big), \quad k = 0, 1, \ldots, N; i = 1, \ldots, N, \quad (4.1)$$

where

$$\widetilde{P}^*_{k,i}(\omega) = 1 - \sum_{j=0}^{k-1}\frac{p_{i-k,i-j}}{1 - p_{i-k,i-k}}\big(1 - \widetilde{B}^*_{j,i}(\omega)\big), \qquad k = 0, 1, \ldots, N; i = 1, \ldots, N. \quad (4.2)$$

Given this modified definition of $\widetilde{B}^*_{k,i}(\omega)$, the function $\widetilde{R}^*_{k,i}(\omega)$ remains unchanged. The expression for the LST of the cycle time C_i, given by (3.12), also remains valid for systems containing exhaustively served queues.

4.3 Waiting-time distributions

Internal customers. The waiting time LST of internal customers (3.14) is determined by conditioning on the event that an arrival in Q_i follows a service completion in some Q_{i-k}. As

stated before, for queues with exhaustive service we need to take into account that customers that are routed back to the same queue will be served during the same visit period. For an arbitrary exhaustively served queue Q_e, this results in

$$\widetilde{W}_e^I(\omega) = \sum_{k=0}^{N-1} \frac{\gamma_{e-k} p_{e-k,e}}{\gamma_e - \lambda_e} \widetilde{WC}_e^{(B_{e-k})}(\omega). \tag{4.3}$$

Compared to (3.14), the summation starts at $k = 0$ and runs up to $k = N - 1$. We now introduce

$$\mathbf{B'_{0,i}} = \big(1, \ldots, 1, \widetilde{B}_i(\omega), 1, \ldots, 1\big), \qquad i = 1, \ldots, N,$$

with $\widetilde{B}_i(\omega)$ at the position corresponding to customers in Q_i. If Q_i has exhaustive service, there is a subtle difference with $\mathbf{B_{0,i}}$ which has $\widetilde{BP}_i(1, \omega)$ at position i. We can now determine $\widetilde{WC}_e^{(B_{e-k})}(\omega)$ for any Q_e that receives exhaustive service:

$$\widetilde{WC}_e^{(B_{e-k})}(\omega) = \widetilde{LC}^{(B_{e-k})}\big(\mathbf{B'_{0,e}} \bigotimes_{j=0}^{k-1} \mathbf{B_{j,e-1}}\big) \prod_{j=0}^{k-1} \widetilde{R}^*_{j,e-1}(\omega), \qquad k = 1, \ldots, N-1,$$

$$\widetilde{WC}_e^{(B_e)}(\omega) = \widetilde{LC}^{(B_e)}\big(\mathbf{B'_{0,e}}\big).$$

For each Q_i that receives gated service, we can still use (3.14) with the modified definition of $\widetilde{B}^*_{k,i}(\omega)$ for each Q_{i-k} which receives exhaustive service.

External customers. The waiting time LST of external customers (3.17) is determined by conditioning on the event that an arrival in Q_i takes place during $V_{i-1}, \ldots, V_{i-N}$ or during $R_{i-1}, \ldots, R_{i-N}$. Before discussing the waiting times of external customers arriving in an exhaustively served queue, it is important to realise that allowing some queues to have exhaustive service will now also require some changes to waiting times of customers arriving in a queue with gated service. This means that (3.23) should now become

$$\begin{aligned}\widetilde{W}_i^{(V_{i-k})}(\omega) = \widetilde{B}^{PR}_{i-k}\Big(&\sum_{j=1}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j-1,i-1}(\omega)\big) + \lambda_i\big(1 - \widetilde{B}_i(\omega)\big) + \lambda_{i-k}\big(1 - \widetilde{B}^*_{k-1,i-1}(\omega)\big), \\ &\omega + \sum_{j=1}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j-1,i-1}(\omega)\big) + \lambda_{i-k}\big(1 - \widetilde{B}^*_{k-1,i-1}(\omega)\big)\Big) \\ \times\, \widetilde{LB}^{(B_{i-k})}\big(\mathbf{B_{0,i}} &\bigotimes_{j=0}^{k-1} \mathbf{B_{j,i-1}}\big) \prod_{j=0}^{k-1} \widetilde{R}^*_{j,i-1}(\omega) \times \frac{1 - \sum_{j=0}^{k-1} p_{i-k,i-j-1}\big(1 - \widetilde{B}^*_{j,i-1}(\omega)\big)}{\widetilde{B}^*_{k-1,i-1}(\omega)},\end{aligned} \tag{4.4}$$

if Q_{i-k} receives exhaustive service (and Q_i receives gated service). Compared to (3.23) we can see that there are two additional terms $\lambda_{i-k}\big(1 - \widetilde{B}^*_{k-1,i-1}(\omega)\big)$ which take into account that customers arriving in Q_{i-k} during the elapsed *and* during the residual part of the present service time B_{i-k} will be served during the present visit period. Furthermore, we can see that $\widetilde{P}^*_{k-1,i-1}(\omega)$ has been replaced by $1 - \sum_{j=0}^{k-1} p_{i-k,i-j-1}\big(1 - \widetilde{B}^*_{j,i-1}(\omega)\big)$, which is required because the customer being served should be allowed to return to Q_{i-k} upon his service completion.

If Q_e receives exhaustive service we have to make some additional changes. We have

$$\widetilde{W}_e^E(\omega) = \frac{1}{\mathbb{E}[C]} \sum_{k=1}^{N} \left(\mathbb{E}[V_{e-k+1}] \widetilde{W}_e^{(V_{e-k+1})}(\omega) + r_{e-k} \widetilde{W}_e^{(R_{e-k})}(\omega) \right), \tag{4.5}$$

where we have chosen to denote the waiting time LST of customers arriving in Q_e during V_e as $\widetilde{W}_e^{(V_e)}(\omega)$ rather than $\widetilde{W}_e^{(V_{e-N})}(\omega)$ to illustrate the fact that they will be served during the same visit period. The expression for $\widetilde{W}_e^{(R_{e-k})}(\omega)$, given by (3.19), should be slightly modified if Q_e receives exhaustive service. However, since the only required modification is that $\mathbf{B}_{\boldsymbol{0},\boldsymbol{i}}$ should be replaced by $\mathbf{B}'_{\boldsymbol{0},\boldsymbol{i}}$, we refrain from giving the complete expression.

If $k > 0$, the expression for $\widetilde{W}_e^{(V_{e-k})}(\omega)$ remains almost the same as (3.23) if Q_{e-k} receives gated service, or (4.4) if Q_{e-k} receives exhaustive service. The only change is, once again, that $\mathbf{B}_{\boldsymbol{0},\boldsymbol{i}}$ should be replaced by $\mathbf{B}'_{\boldsymbol{0},\boldsymbol{i}}$. The case $k = 0$ results in a much simpler expression, since we only have to wait for the service times of the customers that were present at the beginning of the present service (excluding the customer in service) plus the service times of the customers that have arrived in Q_e during the elapsed part of the present service, plus the residual service time:

$$\widetilde{W}_e^{(V_e)}(\omega) = \widetilde{B}_e^{PR}\Big(\lambda_e\big(1 - \widetilde{B}_e(\omega)\big), \omega\Big) \frac{\widetilde{LB}^{(B_e)}\big(\mathbf{B}'_{\boldsymbol{0},\boldsymbol{e}}\big)}{\widetilde{B}_e(\omega)}.$$

Arbitrary customers. The LST of the waiting-time distribution of an arbitrary customer in an exhaustively served queue immediately follows after conditioning on the event that an arbitrary customer is either an internal or an external customer, similar to the derivation of (3.25). The result is presented in the theorem below.

Theorem 4.1 The LST of the waiting-time distribution of an arbitrary customer in Q_i, if this queue receives exhaustive service, is given by:

$$\widetilde{W}_i(\omega) = \frac{\gamma_i - \lambda_i}{\gamma_i} \widetilde{W}_i^I(\omega) + \frac{\lambda_i}{\gamma_i} \widetilde{W}_i^E(\omega), \qquad i = 1, \ldots, N, \tag{4.6}$$

where $\widetilde{W}_i^I(\omega)$ and $\widetilde{W}_i^E(\omega)$ are defined in (4.3) and (4.5).

5 Applicability of the model

In this section we give some numerical examples that indicate the versatility of the model that we have discussed. To this end, we use some examples that can be found in the existing literature, and show how our model can be used to describe the various systems and find the relevant performance measures. Hence, most of the results presented in this section are not novel, but the way of deriving them is new.

Example 1: tandem queues with parallel queues in the first stage. We first use an example that was introduced by Katayama [19], who studies a network consisting of three queues. Customers arrive at Q_1 and Q_2, and are routed to Q_3 after being served (see Figure

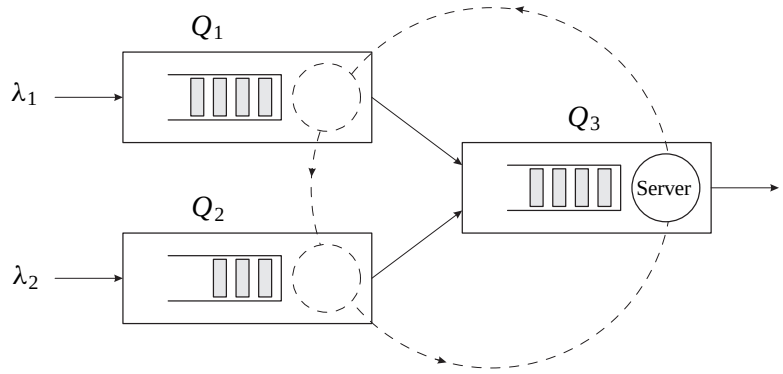

Figure 1: Tandem queues with parallel queues in the first stage, as discussed in Example 1.

1). This model, which is referred to as a tandem queueing model with parallel queues in the first stage, is a special case of the model discussed in the present paper. We simply put $p_{1,3} = p_{2,3} = p_{3,0} = 1$ and all other $p_{i,j}$ are zero. We use the same values as in [19]: $\lambda_1 = \lambda_2/10$, service times are deterministic with $b_1 = b_2 = 1$, and $b_3 = 5$. The server serves the queues exhaustively, in cyclic order: 1, 2, 3, 1, ... The only difference with the model discussed in [19] is that we introduce (deterministic) switch-over times $R_2 = R_3 = 2$. We assume that no time is required to switch between the two queues in the first stage, so $r_1 = 0$. In Figure 2 we show the means and standard deviations of the waiting times of customers at the three queues. These plots reveal that in the heavy-traffic regime, as $\rho \uparrow 1$, the *mean* waiting times of customers in Q_3 are close to those in Q_1, but the *standard deviations* of the waiting times in Q_3 are closer to those in Q_2. Further inspection of the exact results, obtained by differentiating the LSTs, confirms that in both cases the limits are very close, but not exactly the same.

It is also interesting to study the light-traffic behaviour of the system, i.e., as $\rho \downarrow 0$. From the plots in Figure 2 we can see that, as $\rho \downarrow 0$, the *mean* waiting times are all equal, but the *standard deviation* of the waiting times in Q_1 and Q_2 is different than in Q_3. From the LSTs of the waiting-time distributions we can obtain the exact expressions when $\rho \downarrow 0$, by taking the Taylor expansion in ρ at $\rho = 0$ and subsequently ignoring all $\mathcal{O}(\rho)$ terms. This, combined with the fact that $R_1 = 0$ and all of the routing probabilities are either 0 or 1, considerably simplifies all expressions from the previous section:

$$\widetilde{W}_1(\omega) = \widetilde{W}_1^E(\omega) \to \frac{r_2}{r}\widetilde{W}_1^{(R_2)}(\omega) + \frac{r_3}{r}\widetilde{W}_1^{(R_3)},$$
$$\widetilde{W}_2(\omega) = \widetilde{W}_2^E(\omega) \to \frac{r_2}{r}\widetilde{W}_2^{(R_2)}(\omega) + \frac{r_3}{r}\widetilde{W}_2^{(R_3)},$$
$$\widetilde{W}_3(\omega) = \widetilde{W}_3^I(\omega) \to \frac{\lambda_1}{\lambda_1+\lambda_2}\widetilde{WC}_3^{(B_1)}(\omega) + \frac{\lambda_2}{\lambda_1+\lambda_2}\widetilde{WC}_3^{(B_2)}(\omega).$$

Since we are considering the case $\rho \downarrow 0$, these expressions can be simplified even further to closed-form expressions, because ignoring all $\mathcal{O}(\rho)$ terms is equivalent to regarding the system

as being empty all the time:

$$\widetilde{W}_1(\omega) \to \frac{r_2}{r}\widetilde{R}_2^{PR}(0,\omega)\widetilde{R}_3(\omega) + \frac{r_3}{r}\widetilde{R}_3^{PR}(0,\omega),$$
$$\widetilde{W}_2(\omega) \to \frac{r_2}{r}\widetilde{R}_2^{PR}(0,\omega)\widetilde{R}_3(\omega) + \frac{r_3}{r}\widetilde{R}_3^{PR}(0,\omega),$$
$$\widetilde{W}_3(\omega) \to \frac{\lambda_1}{\lambda_1+\lambda_2}\widetilde{R}_1(\omega)\widetilde{R}_2(\omega) + \frac{\lambda_2}{\lambda_1+\lambda_2}\widetilde{R}_2(\omega).$$

These expressions reveal the true behaviour of the system in light traffic. The waiting times in Q_1 and Q_2 are simply the total residual switch-over time, with mean $r^{(2)}/2r = 2$ and second moment $r^{(3)}/3r = 16/3$. For queue Q_3 the situation is different, because this queue only contains internally rerouted customers. Customers being rerouted from Q_1 have to wait for the switch-over times $R_1 + R_2$, whereas customers arriving from Q_2 have to wait only for R_2. Since $R_1 = 0$, the waiting time only consists of $R_2 = 2$ in both cases. Substituting all parameter values results in the following LT limits of the waiting-time LSTs:

$$\widetilde{W}_1(\omega) \to \frac{1-\mathrm{e}^{-4\omega}}{4\omega}, \qquad \widetilde{W}_2(\omega) \to \frac{1-\mathrm{e}^{-4\omega}}{4\omega}, \qquad \widetilde{W}_3(\omega) \to \mathrm{e}^{-2\omega} \qquad (\rho \downarrow 0).$$

Differentiating the LSTs gives the following results as $\rho \downarrow 0$:

$$\mathbb{E}[W_1] \to 2, \qquad \mathbb{E}[W_2] \to 2, \qquad \mathbb{E}[W_3] \to 2,$$
$$\mathrm{sd}[W_1] \to \sqrt{4/3}, \qquad \mathrm{sd}[W_2] \to \sqrt{4/3}, \qquad \mathrm{sd}[W_3] \to 0.$$

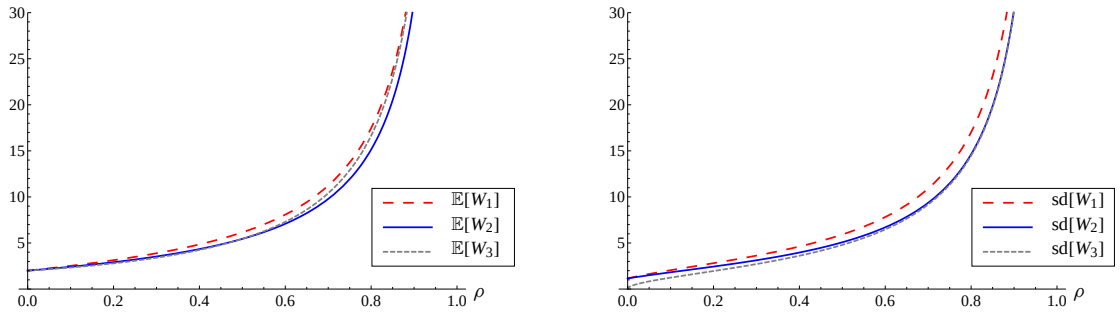

Figure 2: Means and standard deviations of the waiting times in the first numerical example.

Example 2: a two-stage queueing model with customer feedback. This second example is introduced by Takács [30], and extended by Ali and Neuts [1]. The queueing system under consideration consists of a waiting room, in which customers arrive according to a Poisson process with intensity λ, and a service room. The customers are all transferred simultaneously to the service room where they receive service in order of arrival. However, at the moment of the transfer to this service room M additional "overhead customers" are added to the front of this queue. (In [30] M is a constant, in [1] it is a random variable.) Upon service completion, each customer leaves the system with probability q, and returns to the waiting room with probability $1 - q$. Overhead customers leave the system with probability

one after being served. As soon as the last customer in the service room finishes service (and either leaves the system, or returns to the waiting room) all customers present in the waiting room are transferred to the service room, where they will receive service after a new batch of overhead customers has been served, and so on. A schematic representation of this model is depicted in Figure 3.

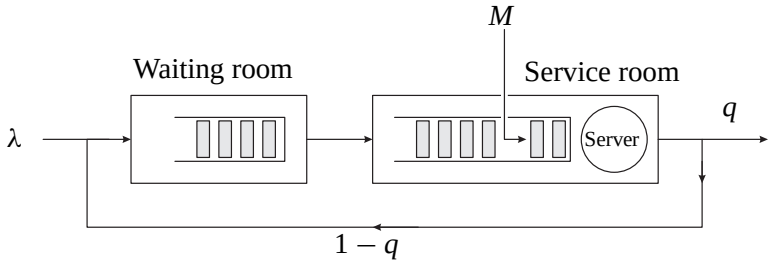

Figure 3: The two-stage queueing model with customer feedback, as discussed in Example 2.

We use the same input parameters as Takács [30]: $q = 2/3$ and $\lambda/\mu = 1/6$, where $1/\mu$ is the mean service time in the service room. This service time is exponentially distributed. The number of overhead customers that are added to the front of the queue is a constant with value M. We can model this system in terms of our network with a single, shared server by defining arrival intensities $\lambda_1 = \lambda$ and $\lambda_2 = 0$. The service times in stations 1 and 2 are respectively 0 and exponentially distributed with mean $b_2 = 1/\mu$. The routing probabilities are $p_{1,2} = 1$ and $p_{2,1} = 1/3$, the other $p_{i,j}$ are zero. The service times of the overhead customers are also exponentially distributed with parameter μ. Hence, we can model the addition of M overhead customers as a switch-over time which is Erlang-M distributed with parameter μ. The switch-over time between Q_2 and Q_1 is zero. Note that, since $b_1 = 0$, there is no difference between gated and exhaustive service. By differentiation of the waiting time LSTs (3.25), we can obtain explicit expressions for all moments of the waiting-time distributions for this example. The first three moments of the waiting times are given below.

$$\mathbb{E}[W_1] = \frac{1+M}{2\mu}, \quad \mathbb{E}[W_1^2] = \frac{(M+1)(11M+25)}{27\mu^2}, \quad \mathbb{E}[W_1^3] = \frac{(M+1)(M(43M+223)+310)}{108\mu^3},$$
$$\mathbb{E}[W_2] = \frac{1+7M}{6\mu}, \quad \mathbb{E}[W_2^2] = \frac{(M+1)(37M+11)}{27\mu^2}, \quad \mathbb{E}[W_2^3] = \frac{(M+1)(M+2)(175M+81)}{108\mu^3}.$$

The results are slightly different from those presented in [30], because Takács also considers the overhead customers in the computations of the waiting times and allows them to return to the waiting room after their service is completed. Modelling this situation would require one minor adaptation in the laws of motion (adding the overhead customers at the beginning of V_2) and another adaptation in the waiting time LST (conditioning on the event that a new customer is an overhead customer). These changes are not too difficult but beyond the scope of this paper.

6 Discussion and further research

In this section, we not only elaborate on the developed method and its applicability, but we also discuss possible ways of extending the present study.

Method. As mentioned in the introduction, the main complicating factor of the model under consideration is caused by the rerouting of internal customers. This implies that the *total* arrival process at each queue is not Poisson, and not even renewal. Traditional methods to determine waiting-time distributions in each queue are based on the distributional form of Little's Law, which relies on the assumption of Poisson arrivals. Contrary to the distributional form of Little's Law, we explicitly make use of the branching structure to find waiting-time distributions. The main idea is that upon the arrival of a tagged customer Y at time t at Q_i we compute *a priori* the total future service times at each of the queues, for *all* the other customers present in the system at time t that will be served before customer Y enters service at Q_i (see (3.7)). Additionally, we add the total future service requirements of all external arrivals (and their descendants) that will be served before customer Y enters service (see (3.9)). The advantage of this method is that a system no longer needs to satisfy all of the prerequisites required to apply the distributional form of Little's Law (see [21]).

Applicability. The novel approach of this paper to find the LST of the waiting-time distribution can also be applied to other types of models with a single server serving multiple queues. Obviously, one can apply it to standard polling models (without customer routing) by simply taking $p_{i,0} = 1$ and $p_{i,j} = 0$ for $j > 0$. However, the developed methodology carries almost directly over to tandem queues [23, 34], multi-stage queueing models with parallel queues [19], feedback vacation queues [9, 33], symmetric feedback polling systems [31, 33], systems with a waiting room [1, 30], closed networks [2], $M/G/1$ queues with permanent and transient customers [8], networks with permanent and transient customers [3], or polling models with arrival rates that depend on the location of the server [4, 7].

Further research. Since the model can be described as a multi-type branching process, *explicit closed-form* expressions can be obtained for the waiting-time distributions under heavy-traffic (HT) assumptions. Such expressions are appealing because they give fundamental insight in how the system performance depends on the system parameters, and in particular on the routing probabilities $p_{i,j}$. HT asymptotics can be obtained by combining insights from multi-type branching processes [35], fluid analyses [24, 25], and the heavy-traffic averaging principle by Coffman et al. [12, 13]. The HT analysis is relevant because in practice the proper operation of the system is particularly important when the system is heavily loaded. The HT asymptotics form an excellent basis for the development of approximations for the waiting-time distributions for *arbitrary* loads. For the *mean* waiting times, preliminary results are obtained in [6].

From a practical perspective, motivated by applications in production systems [5], an important extension of the model under consideration is a model where customers visit a predetermined, class-specific sequence of queues in a fixed order. In our model one would have to define multiple customer classes, each having their own fixed visit order through the system. The method presented in this paper forms a good basis for this extension.

Acknowledgements

The authors are grateful to Ivo Adan and Onno Boxma for providing valuable comments on earlier drafts of the present paper.

References

[1] O. M. E. Ali and M. F. Neuts. A service system with two stages of waiting and feedback of customers. *Journal of Applied Probability*, 21:404–413, 1984.

[2] E. Altman and U. Yechiali. Polling in a closed network. *Probability in the Engineering and Informational Sciences*, 8(3):327–343, 1994.

[3] R. Armony and U. Yechiali. Polling systems with permanent and transient jobs. *Communications in Statistics. Stochastic Models*, 15(3):395–427, 1999.

[4] M. A. A. Boon, A. C. C. van Wijk, I. J. B. F. Adan, and O. J. Boxma. A polling model with smart customers. *Queueing Systems*, 66(3):239–274, 2010.

[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*, 16:67–82, 2011.

[6] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Queueing networks with a single shared server: light and heavy traffic. *SIGMETRICS Performance Evaluation Review*, 39(2):44–46, 2011.

[7] O. J. Boxma. Polling systems. In K. Apt, L. Schrijver, and N. Temme, editors, *From universal morphisms to megabytes: A Baayen space odyssey – Liber amicorum for P. C. Baayen*, pages 215–230. CWI, Amsterdam, 1994.

[8] O. J. Boxma and J. W. Cohen. The $M/G/1$ queue with permanent customers. *IEEE Journal on Selected Areas in Communications*, 9(2):179–184, 1991.

[9] O. J. Boxma and U. Yechiali. An $M/G/1$ queue with multiple types of feedback and gated vacations. *Journal of Applied Probability*, 34:773–784, 1997.

[10] O. J. Boxma, J. Bruin, and B. H. Fralix. Waiting times in polling systems with various service disciplines. *Performance Evaluation*, 66:621–639, 2009.

[11] O. J. Boxma, O. Kella, and K. M. Kosiński. Queue lengths and workloads in polling systems. *Operations Research Letters*, 39:401–405, 2011.

[12] E. G. Coffman, Jr., A. A. Puhalskii, and M. I. Reiman. Polling systems with zero switchover times: A heavy-traffic averaging principle. *The Annals of Applied Probability*, 5(3):681–719, 1995.

[13] E. G. Coffman, Jr., A. A. Puhalskii, and M. I. Reiman. Polling systems in heavy-traffic: A Bessel process limit. *Mathematics of Operations Research*, 23:257–304, 1998.

[14] M. Eisenberg. Queues with periodic service and changeover time. *Operations Research*, 20(2):440–451, 1972.

[15] S. Foss. Queues with customers of several types. In A. A. Borovkov, editor, *Advances in Probability Theory: Limit Theorems and Related Problems*, pages 348–377. Optimization Software, 1984.

[16] Y. Gong and R. de Koster. A polling-based dynamic order picking system for online retailers. *IIE Transactions*, 40:1070–1082, 2008.

[17] S. E. Grasman, T. L. Olsen, and J. R. Birge. Setting basestock levels in multiproduct systems with setups and random yield. *IIE Transactions*, 40(12):1158–1170, 2008.

[18] D. Grillo. Polling mechanism models in communication systems – some application examples. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 659–699. North-Holland, Amsterdam, 1990.

[19] T. Katayama. A cyclic service tandem queueing model with parallel queues in the first stage. *Stochastic Models*, 4:421–443, 1988.

[20] V. Kavitha and E. Altman. Queueing in space: design of message ferry routes in static adhoc networks. In *Proceedings ITC21*, 2009.

[21] J. Keilson and L. D. Servi. The distributional form of Little's Law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9(4):239–247, 1990.

[22] H. Levy and M. Sidi. Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communications*, 38:1750–1760, 1990.

[23] S. S. Nair. A single server tandem queue. *Journal of Applied Probability*, 8(1):95–109, 1971.

[24] T. L. Olsen and R. D. van der Mei. Polling systems with periodic server routeing in heavy traffic: distribution of the delay. *Journal of Applied Probability*, 40:305–326, 2003.

[25] T. L. Olsen and R. D. van der Mei. Periodic polling systems in heavy-traffic: renewal arrivals. *Operations Research Letters*, 33:17–25, 2005.

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

[27] D. Sarkar and W. I. Zangwill. File and work transfers in cyclic queue systems. *Management Science*, 38(10):1510–1523, 1992.

[28] M. Sidi and H. Levy. Customer routing in polling systems. In P. King, I. Mitrani, and R. Pooley, editors, *Proceedings Performance '90*, pages 319–331. North-Holland, Amsterdam, 1990.

[29] M. Sidi, H. Levy, and S. W. Fuhrmann. A queueing network with a single cyclically roving server. *Queueing Systems*, 11:121–144, 1992.

[30] L. Takács. A queuing model with feedback. *Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, 11(4):345–354, 1977.

[31] H. Takagi. Analysis and applications of a multiqueue cyclic service system with feedback. *IEEE Transactions on Communications - TCOM*, 35(2):248–250, 1987.

[32] H. Takagi. Analysis and application of polling models. In G. Haring, C. Lindemann, and M. Reiser, editors, *Performance Evaluation: Origins and Directions*, volume 1769 of *Lecture Notes in Computer Science*, pages 424–442. Springer Verlag, Berlin, 2000.

[33] T. Takine, H. Takagi, and T. Hasegawa. Sojourn times in vacation and polling systems with Bernoulli feedback. *Journal of Applied Probability*, 28(2):422–432, 1991.

[34] M. Taube-Netto. Two queues in tandem attended by a single server. *Operations Research*, 25(1):140–147, 1977.

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