EURANDOM PREPRINT SERIES
2011-044

Waiting times in queueing networks with a single shared server

M.A.A. Boon, R.D. van der Mei, E.M.M. Winands

ISSN 1389-2355

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 [‡]
emm.winands@few.vu.nl

December 19, 2011

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.

The present research develops a novel unifying framework to find the waiting time distribution, which can be applied to a wide variety of models which lacked an analysis of the waiting time distribution until now. That is, we derive the waiting time distributions for stable systems as well as various asymptotic results (heavy traffic, light traffic, and infinite switch-over times) for systems with general renewal arrival processes. By interpolating between these asymptotic regimes, we develop simple closed-form approximations for the waiting time distribution for arbitrary loads.

Keywords: queueing network, waiting times, heavy traffic, light traffic, approximation

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 probability $p_{i,j}$, or leaves the system with probability $p_{i,0}$. This model can be seen as an extension of the standard polling model (in which

[*] The research was done in the framework of the BSIK/BRICKS project, and of the European Network of Excellence Euro-NF.

[†] 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

[§] Centre for Mathematics and Computer Science (CWI), 1098 SJ Amsterdam, The Netherlands 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 goal of the present paper is the derivation of the waiting time distribution in a queueing network with a single shared server.* In most of the paper we assume that each queue receives gated service (only those customers present at the server's arrival at a queue will be served before the server switches to the next queue). The analysis of systems with gated service is slightly more involved than systems with exhaustive service. For completeness, we discuss the case where (some of) the queues receive exhaustive service in the appendix.

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, 19, 23, 33] 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 [18], 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 [21], a dynamic order picking system where the order picker drops off the picked items at the depot where sorting of the items is performed [17], and an internal mail delivery system where a clerk continuously makes rounds within the offices to pick up, sort and deliver mail [28].

The key observation, which is at the same time the mathematical motivation of the present study, is the fact that internally rerouted customers do not arrive at queues according to standard Poisson processes. The standard school of deriving delay distributions is, however, the one embroidering the distributional form of Little's Law, which relies heavily on the assumption that every customer in the system has arrived according to a Poisson process. Due to this intrinsic complexity of the model, studies in the past were restricted to queue lengths and *mean* delay figures (see [6, 28, 29, 30]). This motivates us to develop a novel framework to derive the waiting time *distribution* - a performance metric of which the importance requires no further explanation.

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 [33], tandem queues [24, 35], multi-stage queueing models with parallel queues [20], feedback vacation queues [10, 34], symmetric feedback polling systems [32, 34], systems with a waiting room [1, 31], 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 the present paper is twofold. Firstly, we derive the Laplace-Stieltjes transform (LST) of the waiting time distribution of an arbitrary (internally rerouted, or external) customer in a queueing network with a single shared server. Although the *mean* waiting times have already been studied in the past, no results have been known in the existing literature for the waiting time *distribution*. Since the interdependence of the queueing processes prohibits an exact *explicit* analysis with closed-form expressions, we also derive various asymptotic (heavy traffic, light traffic and infinite switch-over times) and approximate results for the waiting time distribution in systems with general renewal arrival processes. These closed-form expressions are strikingly simple and show explicitly how the delays depend on the system parameters and in particular on the routing probabilities $p_{i,j}$. Numerical results are presented to assess the accuracy of the distributional approximation.

Secondly, a novel method is developed to find the waiting time distribution in queueing systems, which can be applied to a myriad of models which lacked an analysis of the waiting time distribution until now. Contrary to existing methods, we explicitly make use of the branching structure to find waiting time distributions. 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. That is, 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, 10, 20, 24, 31, 32, 33, 34, 35] but also, for example, in a closed network [2], in an $M/G/1$ queue with permanent and transient customers [9], 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, 8]. 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 general loads. The penultimate section, where we relax the assumption of Poisson arrivals, studies the behaviour of our system under heavy-traffic conditions. In the last section we derive an accurate closed-form approximation of the waiting time distribution based on asymptotic results, and we present some examples which show the wide range of applicability of the studied model. Systems with a mixture of gated and exhaustive service are discussed in the appendix.

2 Model description and notation

We consider a queueing network consisting of $N \geq 1$ infinite buffer queues $Q_1, \dots, Q_N$. 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 $b_i := \mathbb{E}[B_i]$ and LST $\widetilde{B}_i(\cdot)$. In general we denote the LST or 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 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. 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. 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 $\sum_{j=0}^{N} p_{i,j} = 1$ for all i, and that the transition of a customer from Q_i to Q_j takes no time. The model under consideration has a branching structure, which is discussed in more detail by Resing [27]. The total arrival rate at Q_i is denoted by γ_i, which is the unique solution of the following set of linear equations:

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

The offered load to Q_i is $\rho_i := \gamma_i b_i$ and the total utilisation 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 [30]). The total service time B_i^* of a customer is the total amount of service given during the presence of the customer in the network. Its first moment, denoted by β_i, is uniquely determined by the following set of linear equations: For $i = 1, \dots, N$,

$$\beta_i = b_i + \sum_{j=1}^{N} \beta_j p_{i,j}.$$

The LST of B_i^* is not discussed in the present paper, but can be obtained by solving a similar set of equations.

3 The waiting time distribution

In the present section we study the waiting time distribution of an arbitrary customer. We define the waiting time W_i as the time between a customer's arrival at Q_i 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 distribution at embedded epochs

Sidi et al. [30] 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 $\mathbf{z} = (z_1, \dots, z_N, z_G)$. The element z_i, $i = 1, \dots, N$, 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 appendix). Before studying the joint queue length distributions, we briefly introduce some convenient notation:

$$\Sigma(\mathbf{z}) = \sum_{j=1}^{N} \lambda_j (1 - z_j),$$

$$\Sigma_i(\mathbf{z}) = \lambda_i (1 - z_G) + \sum_{j \neq i} \lambda_j (1 - z_j),$$

$$P_i(\mathbf{z}) = p_{i,0} + p_{i,i} z_G + \sum_{j \neq i} p_{i,j} z_j.$$

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 $i = 1, \dots, N$. A cycle C_i starts with a visit to Q_i and consists of the periods V_i, R_i, $V_{i+1}, \dots, V_{i+N-1}$, 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 [27]), we can express each of these functions in terms of $\widetilde{LB}^{(V_i)}(\mathbf{z})$, for any $i = 1, \dots, N$. These relations, which are sometimes called *laws of motion*, are given below.

$$\begin{aligned}
\widetilde{LC}^{(V_i)}(\mathbf{z}) &= \widetilde{LB}^{(V_i)}\Big(z_1, \dots, z_{i-1}, \widetilde{B}_i\big(\Sigma_i(\mathbf{z})\big)P_i(\mathbf{z}), z_{i+1}, \dots, z_N, z_G\Big), \qquad (3.1)\\
\widetilde{LB}^{(R_i)}(\mathbf{z}) &= \widetilde{LC}^{(V_i)}(z_1, \dots, z_N, z_i),\\
\widetilde{LC}^{(R_i)}(\mathbf{z}) &= \widetilde{LB}^{(R_i)}(\mathbf{z})\widetilde{R}_i\Big(\Sigma(\mathbf{z})\Big),\\
\widetilde{LB}^{(V_{i+1})}(\mathbf{z}) &= \widetilde{LC}^{(R_i)}(\mathbf{z}),\\
&\vdots\\
\widetilde{LB}^{(V_{i+N})}(\mathbf{z}) &= \widetilde{LC}^{(R_{i+N-1})}(\mathbf{z}). \qquad (3.2)
\end{aligned}$$

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 [27] 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 (3.1)-(3.2) can be differentiated with respect to the variables $z_1, \dots, z_N, z_G$. 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 [16] has observed the following relation, albeit in a slightly different model:

$$\widetilde{LB}^{(V_i)}(\mathbf{z}) + \gamma_i \mathbb{E}[C]\widetilde{LC}^{(B_i)}(\mathbf{z})P_i(\mathbf{z}) = \widetilde{LC}^{(V_i)}(\mathbf{z}) + \gamma_i \mathbb{E}[C]\widetilde{LB}^{(B_i)}(\mathbf{z}). \qquad (3.3)$$

Equation (3.3) 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:

$$\widetilde{LC}^{(B_i)}(\mathbf{z}) = \widetilde{LB}^{(B_i)}(\mathbf{z})\widetilde{B}_i\big(\Sigma_i(\mathbf{z})\big)/z_i. \qquad (3.4)$$

Substitution of (3.4) in (3.3) 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 $(V_1, R_1, \dots, V_N, R_N)$.

$$\widetilde{L}(\mathbf{z}) = \frac{1}{\mathbb{E}[C]} \sum_{j=1}^{N} \left(\mathbb{E}[V_j]\widetilde{L}^{(V_j)}(\mathbf{z}) + r_j \widetilde{L}^{(R_j)}(\mathbf{z})\right), \tag{3.5}$$

with $\mathbb{E}[V_j] = \rho_j \mathbb{E}[C]$. In (3.5) 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:

$$\widetilde{L}^{(V_j)}(\mathbf{z}) = \widetilde{LB}^{(B_j)}(\mathbf{z}) \frac{1 - \widetilde{B}_j\left(\Sigma_j(\mathbf{z})\right)}{b_j \Sigma_j(\mathbf{z})}, \tag{3.6}$$

$$\widetilde{L}^{(R_j)}(\mathbf{z}) = \widetilde{LB}^{(R_j)}(\mathbf{z}) \frac{1 - \widetilde{R}_j\left(\Sigma(\mathbf{z})\right)}{r_j \Sigma(\mathbf{z})}. \tag{3.7}$$

The interpretation of (3.6) and (3.7) 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. [12].

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, $i = 1, \dots, N$. 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 $R_i, \dots, R_{i+N-1}$, 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. [11]. 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 $B^*_{k,i}$, $i = 1, \dots, N; k = 0, \dots, N$, as the service time of a type $i - k$ (which is understood as $N + i - k$ if $i \leq k$) customer, plus the service times of all of his descendants that will be served before or during the next visit to Q_i. The special case $B^*_{0,i}$ is simply the service time of a type i customer, $i = 1, \dots, N$. A formal definition in terms of LSTs is given below:

$$\widetilde{B}^*_{k,i}(\omega) = \widetilde{B}_{i-k}\Big(\omega + \sum_{j=0}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j,i}(\omega)\big)\Big)\widetilde{P}^*_{k,i}(\omega), \qquad k = 0, \dots, N; i = 1, \dots, N, \tag{3.8}$$

where

$$\widetilde{P}^*_{k,i}(\omega) = 1 - \sum_{j=0}^{k-1} p_{i-k,i-j}\big(1 - \widetilde{B}^*_{j,i}(\omega)\big), \qquad k = 0, \dots, N; i = 1, \dots, N. \tag{3.9}$$

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:

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

Note that, compared to (3.8), 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:

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

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

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.

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

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

Proof:
To prove Theorem 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 $n_1, \dots, n_N$. 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 $R_1, \dots, R_N$,
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 $j = 1, \dots, N$, 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:

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

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) $Q_{j+1}, \dots, Q_{i-1}$ 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 $k = 1, \dots, N$. 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 $R_1, \dots, R_N$ and $S_1, \dots, S_N$. 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 $n_1, \dots, n_N$ for the event that the numbers of customers at the beginning of C_i in queues $1, \dots, N$ are respectively $n_1, \dots, n_N$, we can find the cycle time LST conditional on this event.

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

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 (3.1)-(3.2) to find an expression for $\widetilde{LB}^{(V_i)'}(z_1^E, z_1^I, \dots, z_N^E, z_N^I)$. Just like $\widetilde{LB}^{(V_i)}(z_1, \dots, z_N, z_G)$, $\widetilde{LB}^{(V_i)'}(z_1^E, z_1^I, \dots, z_N^E, z_N^I)$ 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 [22]) to the *external* customers in Q_i:

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

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. 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 $WC_i^{(B_{i-k})}$ $(i, k = 1, \dots, N)$. 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:

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

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 $i, k = 1, \dots, N$.

Theorem 3.3

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

$$\widetilde{WC}_i^{(B_{i-N})}(\omega) = \widetilde{LC}^{(B_i)}\big(\mathbf{B}_{\boldsymbol{N},\boldsymbol{i}} \bigotimes_{j=0}^{N-1} \mathbf{B}_{\boldsymbol{j},\boldsymbol{i-1}}\big) \prod_{j=0}^{N-1} \widetilde{R}^*_{j,i-1}(\omega), \tag{3.16}$$

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

Proof:
The key observation in the proof of Theorem 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 Theorem 3.1, we can express $WC_i^{(B_j)'}$ in terms of $R_1, \ldots, R_N$ and $S_1, \ldots, S_N$:

$$WC_i^{(B_j)'} = \sum_{k=j}^{i-1} \left[\sum_{l=1}^{n_k} S_{k_l} + R_k + \sum_{l=k+1}^{i-1} \sum_{m=1}^{N_l(R_k)} S_{l_m} \right] + \sum_{l=1}^{n_{iG}} B_{i,l}. \tag{3.17}$$

Taking the LST of (3.17) leads to (3.15) if $k < N$, and to (3.16) if $k = N$, after deconditioning. The derivation proceeds along the exact same lines as in the proof of Theorem 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 $W_i^{(R_{i-k})}$ $(i, k = 1, \dots, N)$. 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 $\widetilde{W}_i^{(V_{i-k})}(\cdot)$:

$$\widetilde{W}_i^E(\omega) = \frac{1}{\mathbb{E}[C]} \sum_{k=1}^{N} \left(\mathbb{E}[V_{i-k}] \widetilde{W}_i^{(V_{i-k})}(\omega) + r_{i-k} \widetilde{W}_i^{(R_{i-k})}(\omega) \right), \qquad i = 1, \dots, N. \tag{3.18}$$

We first focus on the waiting time of customers arriving during a switch-over time. Consider a tagged customer arriving in Q_i during $R_{i-k}, i, k = 1, \dots, N$. 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^R_{i-k}. The PGF of the joint queue length distribution at the arrival epoch is given by (3.7). Equation (3.7) 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^P_{i-k}. These random variables are independent. Hence, it is straightforward to adapt (3.7) to find the joint distribution of the queue lengths *and* residual part of R_{i-k}, using the following result from elementary renewal theory:

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

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,

$$\widetilde{L}^{(R_j)}(\mathbf{z}, \omega) = \widetilde{LB}^{(R_j)}(\mathbf{z}) \widetilde{R}_j^{PR}(\Sigma(\mathbf{z}), \omega), \tag{3.19}$$

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.

Theorem 3.4

$$\widetilde{W}_i^{(R_{i-k})}(\omega) = \widetilde{R}^{PR}_{i-k}\Big(\sum_{j=1}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j-1,i-1}(\omega)\big) + \lambda_i\big(1 - \widetilde{B}_i(\omega)\big), \omega + \sum_{j=1}^{k-1} \lambda_{i-j}\big(1 - \widetilde{B}^*_{j-1,i-1}(\omega)\big)\Big)$$
$$\times \widetilde{LB}^{(R_{i-k})}\big(\boldsymbol{B_{0,i}} \bigotimes_{j=0}^{k-2} \boldsymbol{B_{j,i-1}}\big) \prod_{j=0}^{k-2} \widetilde{R}^*_{j,i-1}(\omega), \qquad i, k = 1, \dots, N, \tag{3.20}$$

Proof:
We consider an arbitrary customer arriving in Q_i during R_j. Similar to the proofs of the preceding theorems in this section, we condition on the number of customers present in all queues at the arrival epoch, denoted by $n_1, \dots, n_N$. 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 $n_{j+1}, \dots, n_{i-1}$ customers present at the arrival epoch in $Q_{j+1}, \dots, Q_{i-1}$,
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 $R_{j+1}, \dots, R_{i-1}$,
6. the service of all customers arriving during $t_R, R_{j+1}, \dots, R_{i-1}$ 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* $n_1, \dots, n_N$ and t_R, by $W_i^{(R_j)'}$, we can summarise these items in the following formula:

$$W_i^{(R_j)'} = \sum_{k=j+1}^{i-1}\left[\sum_{l=1}^{n_k} S_{k_l} + R_k + \sum_{l=k+1}^{i-1}\sum_{m=1}^{N_l(R_k)} S_{l_m}\right] + \sum_{l=1}^{n_i} B_{i_l} + t_R + \sum_{l=j+1}^{i-1}\sum_{m=1}^{N_l(t_R)} S_{l_m}. \tag{3.21}$$

Taking the LST of (3.21) and using (3.19) leads to (3.20) after deconditioning. The derivation is not completely straightforward, but rather than providing it here, we refer to the proof of Theorem 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 $i, k = 1, \dots, N$, 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

$$\widetilde{B}_j^{PR}(\omega_P, \omega_R) = \frac{\widetilde{B}_j(\omega_P) - \widetilde{B}_j(\omega_R)}{(\omega_R - \omega_P)b_j}, \qquad j = 1, \dots, N. \tag{3.22}$$

We can now use Equations (3.6) and (3.22) 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:

$$\widetilde{L}^{(V_j)}(\mathbf{z}, \omega) = \widetilde{LB}^{(B_j)}(\mathbf{z})\widetilde{B}_j^{PR}(\Sigma_j(\mathbf{z}), \omega). \tag{3.23}$$

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

Theorem 3.5

$$\widetilde{W}_i^{(V_{i-k})}(\omega) = \widetilde{B}_{i-k}^{PR}\Big(\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big) + \lambda_i\big(1-\widetilde{B}_i(\omega)\big), \omega + \sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)\Big)$$
$$\times \widetilde{LB}^{(B_{i-k})}\Big(\mathbf{B}_{\mathbf{0},i}\bigotimes_{j=0}^{k-1}\mathbf{B}_{\mathbf{j},\mathbf{i-1}}\Big)\prod_{j=0}^{k-1}\widetilde{R}^*_{j,i-1}(\omega)\times\frac{\widetilde{P}^*_{k-1,i-1}(\omega)}{\widetilde{B}^*_{k-1,i-1}(\omega)}, \qquad i=1,\ldots,N; k=1,\ldots,N-1, \tag{3.24}$$

$$\widetilde{W}_i^{(V_{i-N})}(\omega) = \widetilde{B}_{i}^{PR}\Big(\sum_{j=1}^{N-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big) + \lambda_i\big(1-\widetilde{B}_i(\omega)\big), \omega + \sum_{j=1}^{N-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)\Big)$$
$$\times \widetilde{LB}^{(B_{i})}\Big(\mathbf{B}_{N,i}\bigotimes_{j=0}^{N-1}\mathbf{B}_{\mathbf{j},\mathbf{i-1}}\Big)\prod_{j=0}^{N-1}\widetilde{R}^*_{j,i-1}(\omega)\times\frac{\widetilde{P}^*_{N-1,i-1}(\omega)}{\widetilde{B}^*_{N-1,i-1}(\omega)}, \qquad i=1,\ldots,N. \tag{3.25}$$

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:

$$\begin{aligned} W_i^{(V_j)'} &= \sum_{l=1}^{n_j-1} S_{j,l} + \sum_{k=j+1}^{i-1}\sum_{l=1}^{n_k} S_{k_l} + \sum_{l=1}^{n_{iG}} B_{i_l} + \sum_{k=j}^{i-1}\left[R_k + \sum_{l=k+1}^{i-1}\sum_{m=1}^{N_l(R_k)} S_{l_m}\right] \\ &\quad + t_R + \sum_{l=j+1}^{i-1}\sum_{m=1}^{N_l(t_R)} S_{l_m} + \begin{cases} S_l & \text{for } l = j+1,\dots,i-1, \text{ w.p. } p_{j,l}, \\ 0 & \text{w.p. } 1-\sum_{l=j+1}^{i-1} p_{j,l}, \end{cases}. \end{aligned} \tag{3.26}$$

We now show that Equations (3.24) and (3.25) (for the cases $i \neq j$ and $i = j$ respectively) follow from taking the LSTs:

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

where $k = i - j$ (or $k = N + i - j$ if $j \geq i$). Deconditioning of this expression leads to (3.25). □

Arbitrary customers Finally, the LST of the waiting time distribution of an arbitrary customer in Q_i follows from (3.14) and (3.18), after deconditioning on the event that an arbitrary customer is an internal or external customer:

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

Remark 3.6 The novel approach of the present section 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 [24, 35], multi-stage queueing models with parallel queues [20], feedback vacation queues [10, 34], symmetric feedback polling systems [32, 34], systems with a waiting room [1, 31], closed networks [2], $M/G/1$ queues with permanent and transient customers [9], networks with permanent and transient customers [3], or polling models with arrival rates that depend on the location of the server [4, 8].

4 The waiting time distribution under heavy traffic

In the present section we study the behaviour of our system under heavy-traffic (HT) conditions. From now on, we relax the assumption of Poisson arrivals, and we assume that the network consists of at least two stations. We only require that the interarrival times are independent random variables. Heavy-traffic conditions imply that we increase the load of the system until it reaches the point of saturation, $\rho \uparrow 1$. As the total load of the system increases, the visit times, cycle times, and waiting times become larger and will eventually grow to infinity. For this reason, we scale them appropriately and consider the scaled versions. We consider several variables as a function of the load ρ in the system. Scaling is done by varying the interarrival times of the external customers. To be precise, the limit is taken such that the external arrival rates $\lambda_1, \dots, \lambda_N$ are increased, while keeping the service and switch-over time distributions, the routing probabilities and the *ratios* between these arrival rates fixed. For each variable x that is a function of ρ, its value evaluated at $\rho = 1$ is denoted by $\hat{x}$. For $\rho = 1$, the generic interarrival time of the stream in Q_i is denoted by $\hat{A}_i$. Reducing the load ρ is done by scaling the interarrival times, i.e., taking the random variable $A_i := \hat{A}_i/\rho$ as generic interarrival time at Q_i. The (scaled) rate of the arrival stream at Q_i is defined as $\lambda_i = 1/\mathbb{E}[A_i]$. After scaling, the load at Q_i becomes $\rho_i = \rho\hat{\gamma}_i b_i$. Furthermore, we define arrival rates $\hat{\lambda}_i = 1/\mathbb{E}[\hat{A}_i]$, and proportional load at Q_i, $\hat{\rho}_i = \rho_i/\rho$ ("proportional" because $\sum_{i=1}^{N} \hat{\rho}_i = 1$).

To obtain HT-results for the waiting-time distributions, we use HT results for polling systems, which are obtained by Coffman et al. [13, 14] and by Olsen and Van der Mei [25, 26]. The key observation in these papers is the occurrence of a so-called Heavy Traffic Averaging Principle (HTAP). When a polling system becomes saturated, two limiting processes take place. Let V denote the total workload of the system. As the load offered to the system, ρ, tends to 1, the scaled total workload $(1-\rho)V$ tends to a Bessel-type diffusion. However, the work *in each queue* is emptied and refilled at a faster rate than the rate at which the total workload is changing. This implies that during the course of a cycle, the total workload can be considered as constant, while the workloads of the individual queues fluctuate according to a fluid model. The HTAP relates these two limiting processes. We start by discussing the fluid model and subsequently discuss the limiting distribution of the scaled total workload. At the end of this section we use these results to obtain the HT limit of the scaled waiting time distributions.

4.1 Fluid model: workload

We start by studying the fluid limit of the per-queue workload, which is obtained by multiplying by $(1 - \rho)$ and letting $\rho \uparrow 1$. For our model, the fluid limit of the workload at Q_i is a piecewise linear function. During V_k, $k = 1, \dots, N$, *external* fluid particles flow into Q_i at rate $\hat{\lambda}_i$. Each of these fluid

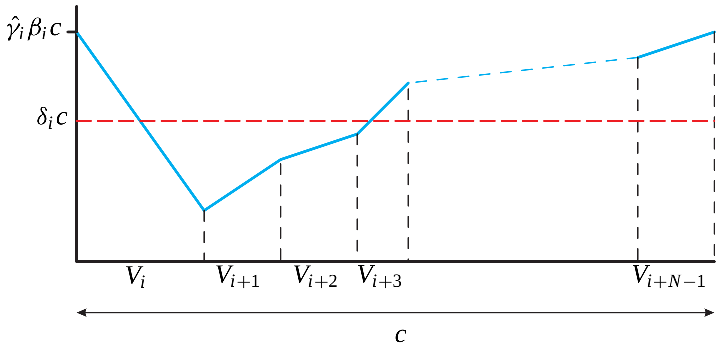


Figure 1: Mean amount of work in Q_i in the fluid limit that arises when the system is in heavy traffic. The length of one cycle is c.

particles brings along β_i units of work into the system. Simultaneously, work is being processed in Q_k at rate one. Since $\sum_{i=1}^{N} \hat{\lambda}_i \beta_i = 1$, the *total* workload remains constant throughout the course of a cycle. Although work is processed at rate one, due to the internal routing work is flowing out of Q_k at rate

$$1 + \frac{1}{b_k} \sum_{i=1}^{N} p_{k,i} \beta_i = \frac{\beta_k}{b_k},$$

which is greater than (or equal to) one. The reason for this anomaly is that work decreases in Q_k either because of the service of fluid particles (customers) in this queue, or because work is shifted due to internal routing of fluid. Work *including* rerouted fluid particles is flowing into Q_i, during V_k, at rate $\hat{\gamma}_{i,k}\beta_i$, where

$$\hat{\gamma}_{i,k} := \hat{\lambda}_i + p_{k,i}/b_k, \qquad i, k = 1, \ldots, N.$$

It is straightforward to verify that $\beta_k/b_k = \sum_{i=1}^{N} \hat{\gamma}_{i,k}\beta_i$. Figure 1 depicts a graphical representation of the mean amount of work in Q_i in the fluid limit throughout the course of a cycle, the length of which is a constant, denoted by c. One can show that the fluid limit of the mean amount of work in Q_i at the beginning of a visit to Q_j is $\sum_{k=i}^{j-1} \hat{\rho}_k \hat{\gamma}_{i,k} \beta_i c$ for $j = i+1, \ldots, i+N$. This reduces to $\hat{\gamma}_i \beta_i c$ for $j = i + N$. We have used that in the fluid limit the fraction of time that the server is visiting Q_j is $\hat{\rho}_j$ $(j = 1, \ldots, N)$. Combining these observations, one can obtain the following expression for δ_i, defined as the ratio of the fluid limit of the average amount of work at Q_i and the length of a cycle (see Figure 1).

Lemma 4.1 For $i = 1, \ldots, N$,

$$\delta_i = \frac{1}{2}\hat{\rho}_i \beta_i (\hat{\gamma}_i + \hat{\rho}_i \hat{\gamma}_{i,i}) + \sum_{j=i+1}^{i+N-1} \hat{\rho}_j \left(\frac{1}{2}\hat{\rho}_j \beta_i \hat{\gamma}_{i,j} + \sum_{k=i}^{j-1} \hat{\rho}_k \beta_i \hat{\gamma}_{i,k} \right). \tag{4.1}$$

As the *total* inflow in all queues is equal to the total outflow per time unit, the total amount of work during a cycle remains constant at level δc, where δ is defined as

$$\delta = \sum_{i=1}^{N} \delta_i. \tag{4.2}$$

4.2 Fluid model: waiting times

For the fluid model under consideration we are interested in the waiting time distribution of an arbitrary fluid particle, internal or external. Just like in the previous section, we define the waiting time as the the time between the arrival in a queue, and the moment of departure from this queue (even if the particle is routed to another, or even the same queue). During V_k fluid flows into Q_i at rate $\hat{\gamma}_{i,k}$. Hence, the probability that an arbitrary fluid particle arrives during V_k, given that it arrives in Q_i, is $\pi_{i,k} := \hat{\gamma}_{i,k}\hat{\rho}_k/\hat{\gamma}_i$. The corresponding waiting time consists of the residual part of V_k, the visit periods $V_{k+1}, \dots, V_{i-1}$, and the processing of the amount of fluid that has arrived in Q_i during the elapsed part of the cycle, i.e., $V_i, \dots, V_{k-1}$ plus the elapsed part of V_k. Let U_k be a uniformly distributed random variable on [0, 1], indicating the fraction of V_k that has elapsed at the arrival epoch of a fluid particle in Q_i. The waiting time distribution is:

$$
\begin{aligned}
W_i^{\mathit{fluid}} &\overset{d}{=} (1-U_k)\hat{\rho}_k c + \sum_{j=k+1}^{i-1} \hat{\rho}_j c + \sum_{j=i-N}^{k-1} \hat{\rho}_j c \hat{\gamma}_{i,j} b_i + U_k \hat{\rho}_k c \hat{\gamma}_{i,k} b_i \qquad \text{w.p. } \pi_{i,k} \\
&= c\Big(1 + \sum_{j=i-N}^{k-1} \hat{\rho}_j (\hat{\gamma}_{i,j} b_i - 1) + U_k \hat{\rho}_k (\hat{\gamma}_{i,k} b_i - 1)\Big) \qquad \text{w.p. } \pi_{i,k},
\end{aligned} \tag{4.3}
$$

for $i = 1, \dots, N$ and $k = i-N, \dots, i-1$.

4.3 Original model: workload, cycle time and waiting times

We now return to the original model under HT conditions. We denote by V the total amount of work in the system at an arbitrary epoch. As far as the total amount of work is concerned, the system behaves like a polling system in heavy traffic with external customers bringing in an amount of work B_i^* in Q_i, but with work shifting from one queue to another upon the service completion of a customer. For polling systems with general renewal arrivals the HT limit of the scaled total amount of work at the beginning of a cycle is conjectured by Olsen and Van der Mei [26]. Although this conjecture is widely accepted to be true, it has only been proven for systems consisting of two queues (cf. [13, 14]), systems with Poisson arrivals (cf. [25]), or for the *means* rather than the complete distributions (cf. [37]). An adaptation of the conjecture in [26] to our model leads to the following result.

Conjecture 4.2 Define

$$
\begin{aligned}
\sigma^2 &= \sum_{i=1}^{N} \hat{\lambda}_i \left(\mathrm{Var}[B_i^*] + (\hat{\lambda}_i \beta_i)^2 \mathrm{Var}[\hat{A}_i]\right), \\
\alpha &= 2r\delta/\sigma^2 + 1, \\
\mu &= 2/\sigma^2,
\end{aligned}
$$

where δ is given by (4.2). Then, for $\rho \uparrow 1$, $(1-\rho)V$ has a Gamma distribution with shape parameter α and rate parameter μ.

For more details we refer to [26] (who, in turn, refer to a result from [14]).

Subsequently, the diffusion limit of the *total* workload process and the workload in the individual queues can be related using the HTAP. To this end, we start with the cycle-time distribution under HT scalings, which follows from Conjecture 4.2 and the fluid analysis carried out in the first part of this section. The length of a cycle depends on the amount of work at the beginning of that cycle (which may be any arbitrarily chosen moment). Denote by $C(x)$ the length of a cycle, given that a total amount of x work is present at its beginning. In steady state, we have the following relation

$$\delta C(x) = x. \tag{4.4}$$

Hence, given an amount of work x, the cycle time is $C(x) = x/\delta$. However, the cycle during which an arbitrary customer arrives, is a so-called *length-biased* cycle. If a random variable X has probability density function $f_X(x)$, then we define the length-biased random variable $\boldsymbol{X}$ as a random variable with probability density function

$$f_{\boldsymbol{X}}(x) = x f_X(x)/\mathbb{E}[X].$$

From renewal theory, we know that the length-biased cycle length accounts for the fact that an arbitrary customer arrives with a higher probability during a long cycle, than during a short one. Hence, when relating the waiting times to the cycle times, one should consider the length-biased cycle time. We are now ready to formulate the second conjecture, concerning the limiting distribution of the scaled length-biased cycle time.

Conjecture 4.3 For $\rho \uparrow 1$, we find that $(1-\rho)\boldsymbol{C_i}$ converges in distribution to a random variable having a Gamma distribution with shape parameter α and rate parameter $\delta\mu$.

Given the cycle time distribution, we can finally find the waiting time distributions under HT conditions. We use the fluid analysis, in combination with the conjectures in this section, to find the limiting distribution of the scaled waiting times. In the fluid analysis the cycle time had a fixed length c. Due to the HTAP we can replace the constant cycle time from the fluid analysis by the random variable $\boldsymbol{C_i}$, the scaled *length-biased* cycle time. Obviously, this replacement can only be carried out because of the independence between the length of the cycle time and the uniformly distributed random variables appearing in (4.3). The following conjecture summarises this result.

Conjecture 4.4 As $\rho \uparrow 1$, the scaled waiting time $(1-\rho)W_i$ converges in distribution to the product of a random variable having the same distribution as W_i^{fluid} and a random variable Γ having the same distribution as the limiting distribution of the scaled length-biased cycle time, $(1-\rho)\boldsymbol{C_i}$. For $i = 1, \dots, N$; $k = i-N, \dots, i-1$, and $\rho \uparrow 1$,

$$(1-\rho)W_i \xrightarrow{d} \Gamma \times \Big(1 + \sum_{j=i-N}^{k-1} \hat{\rho}_j(\hat{\gamma}_{i,j} b_i - 1) + U_k \hat{\rho}_k(\hat{\gamma}_{i,k} b_i - 1)\Big) \text{ w.p. } \pi_{i,k}, \tag{4.5}$$

where Γ is a random variable having a Gamma distribution with parameters α and $\delta\mu$, and $U_1, \dots, U_N$ are independent uniform [0, 1] distributed random variables.

The (HT limit of the) *mean* waiting time of an arbitrary customer in Q_i obviously follows from (4.5), but an easier way to find it, is by application of Little's Law to the mean queue length at Q_i, which is simply the mean amount of work in Q_i divided by the mean total service time.

Corollary 4.5 For $i = 1, \dots, N$,

$$(1-\rho)\mathbb{E}[W_i] \to \left(r + \frac{\sigma^2}{2\delta}\right)\frac{\delta_i}{\hat{\gamma}_i \beta_i}, \qquad (\rho \uparrow 1). \tag{4.6}$$

We conclude this section with some remarks.

Remark 4.6 In the current section we have derived the system behaviour under heavy traffic for systems with general renewal arrival processes based on the partially conjectured HTAP. Recently, Van der Mei [36] has developed a unifying framework to derive rigorous proofs of the heavy-traffic behaviour of branching-type polling models with *Poisson* arrivals. By applying this stepwise approach in conjunction with the results of the previous section to the model under consideration, one can rigorously prove the HT asymptotics in queueing networks served by a single shared server under the assumption of Poisson arrivals. These steps are not particularly enlightening by themselves so we have chosen not to highlight them and refer the interested reader to [36].

Remark 4.7 In HT the system reaches saturation due to an increase in the total utilisation ρ. However, the system might also get saturated due to an increase of the total switch-over time r. These two asymptotic regimes show, however, significantly different behaviour. In [38, 39] it was shown for polling systems that the scaled cycle and intervisit times converge in probability to deterministic quantities in the case that the (deterministic) switch-over times tend to infinity. One has to compare this with the Gamma distribution which is prevalent in the scaled cycle time in the diffusion limit of the present section. The results for polling systems with increasing switch-over times of [38, 39] can be extended to the setting of the current paper. That is, as a consequence of the scaled cycle time converging to a constant, a fluid limit is obtained implying that the scaled delay converges in distribution to a mixture of uniform distributions (cf. Formula (4.3)).

5 Waiting time approximations

The HT diffusion distribution derived in the preceding section may be used directly as an approximation for the waiting time distribution in non-heavy-traffic systems. However, it tends to perform poorly under low or moderate traffic. Therefore, in this section we refine this diffusion distribution such that its mean coincides with the mean of a novel mean waiting time approximation, while the diffusion distribution remains unchanged in the case of HT after refinement (cf. [15]).

5.1 Mean waiting time approximation

In order to derive an approximation for the mean waiting times, we study the LT limit of $\mathbb{E}[W_i]$ which can be found by conditioning on the customer type (external or internally routed).

Theorem 5.1 For $i = 1, \ldots, N$,

$$\mathbb{E}[W_i] \to \frac{\lambda_i}{\gamma_i}\frac{r^{(2)}}{2r} + \sum_{j=i-N}^{i-1} \frac{\gamma_j p_{j,i}}{\gamma_i} \sum_{k=j}^{i-1} r_k, \qquad (\rho \downarrow 0). \tag{5.1}$$

In light traffic we ignore all $O(\rho)$ terms, which implies that we can consider a customer as being alone in the system. Equation (5.1) can be interpreted as follows. An arbitrary customer in Q_i has arrived from outside the network with probability λ_i/γ_i. In this case he has to wait for a residual total switch-over time with mean $r^{(2)}/2r$. If a customer in Q_i arrives after being served in another queue, say Q_j (with probability $\gamma_j p_{j,i}/\gamma_i$), on average he has to wait for the mean switch-over times $r_j, \ldots, r_{i-1}$.

Subsequently, we construct an interpolation between the LT and HT limits that can be used as an approximation for the mean waiting times. For $i = 1, \dots, N$,

$$\mathbb{E}[W_i^{approx}] = \frac{w_i^{LT} + (w_i^{HT} - w_i^{LT})\rho}{1-\rho}, \tag{5.2}$$

where w_i^{LT} and w_i^{HT} are the LT and HT limits respectively, as given in (5.1) and (4.6). Because of the way $\mathbb{E}[W_i^{approx}]$ is constructed, it has the nice properties that it is exact as $\rho \downarrow 0$ and $\rho \uparrow 1$. Furthermore, if we have Poisson arrivals, it satisfies a so-called pseudo-conservation law for the mean waiting times, which is derived in [30]. This implies that the $\mathbb{E}[W_i^{approx}]$ yields exact results for symmetric (and, hence, single-queue) systems.

The astute reader has already noticed that the LT result (5.1) is a first-order Taylor expansion of the mean waiting time at $\rho = 0$, which can be naturally extended with the m^{th} derivatives of the mean waiting time with respect to ρ at $\rho = 0$. Together with the HT limit one has $m + 1$ pieces of information, which can be used to construct an $(m+1)^{th}$ degree polynomial interpolation (cf. [7]). As can be seen in the numerical evaluation, the presented first-order polynomial interpolation is however already quite accurate.

5.2 Refining the HT waiting time distribution

First, let us define $\mathcal{W}_i^{fluid}$ as W_i^{fluid}/c, i.e., the ratio of the waiting time of a particle in the fluid model discussed in the previous section, and the length of a cycle in the fluid model. As a starting point of the refinement of the diffusion distribution, we assume that the waiting time distribution of Q_i for general load can be written as a product of $\mathcal{W}_i^{fluid}$ and a gamma random variable with parameters α_a and μ_{ia}, divided by $(1-\rho)$, in line with the HT result. To parameterise α_a and μ_{ia}, we impose the following three requirements:

1. The refined approximation must coincide with the diffusion distribution (4.5), i.e., $\alpha/\alpha_a \to 1$ and $\mu_i/\mu_{ia} \to 1$ when ρ tends to 1.
2. The mean of the refined approximation equals $\mathbb{E}[W_i^{approx}]$ as defined in (5.2).
3. The squared coefficient of variation of the refined approximation equals the squared coefficient of variation of the HT diffusion distribution (4.5).

These requirements uniquely determine the parameters α_a and μ_{ia}, leading to the following approximation for the waiting time distribution for $\rho < 1$,

$$\mathbb{P}[W_i < x] \approx \mathbb{P}\left[\Gamma_i^{approx} \times \mathcal{W}_i^{fluid} < (1-\rho)x\right], \tag{5.3}$$

where Γ_i^{approx} is a Gamma distributed random variable with parameters

$$\alpha_a = \frac{2r\delta}{\sigma^2} + 1, \text{ and } \mu_{ia} = \frac{\alpha_a \mathbb{E}[\mathcal{W}_i^{fluid}]}{(1-\rho)\mathbb{E}[W_i^{approx}]}. \tag{5.4}$$

It can be shown that this approximation is exact in the limiting case of deterministic set up times that tend to infinity (see [38, 39]) and, by construction, in the HT regime. Finally, it is not inconceivable that the approximation can be refined even further, but since the primary goal of this paper has been the derivation of the waiting time distributions under general and heavy traffic conditions such refinements are beyond the scope of the paper.

5.3 Numerical evaluation

We do not aim at giving an extensive numerical study to assess the accuracy of the approximation. Instead, we give some numerical examples that indicate the versatility of the model that we have discussed, and show the practical usage of the approximation (5.3). 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. It is noteworthy that all of these examples contain one or more queues with exhaustive service, which is described in the appendix.

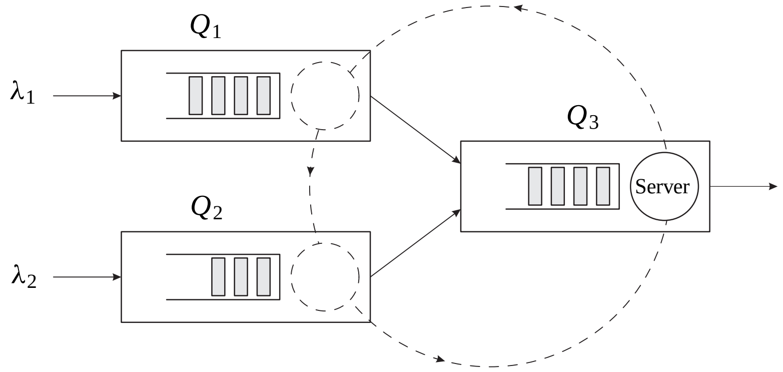


Figure 2: Tandem queues with parallel queues in the first stage, as discussed in Example 1.

Example 1: tandem queues with parallel queues in the first stage. We first use an example that was introduced by Katayama [20], 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 2). 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 [20]: $\lambda_1 = \lambda_2/10$, 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 [20] 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 Table 1 we show the means and standard deviations of the waiting times of customers at the three queues and their approximated values. From this table we can see that the accuracy for the mean waiting time is best for values of ρ close to 0 or 1, but the overall accuracy is very good in general. The standard deviation is approximated very accurately as well, but (in contrast to the mean) its approximation is not exact for the limiting case $\rho \downarrow 0$. Hence, for practical purposes we recommend using it for systems with $\rho > 0.5$.

We have also tested the accuracy of the approximation for different interarrival-time distributions, with squared coefficient of variation (SCV) equal to respectively $\frac{1}{2}$ and 2. In the first case we have fitted a mixed Erlang distribution, and in the second case a hyperexponential distribution. For an SCV equal to $\frac{1}{2}$, the accuracy of the approximations for $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ remains excellent (maximum relative error below 7%). However, for the mean waiting times in Q_3 the performance of the approximation deteriorates, with relative errors up to 30% (for $\rho = 0.5$). The results for an SCV equal to 2 are excellent for all three queues, with maximum relative errors of respectively 5%, 2% and 10%. The accuracy of the approximations for the standard deviations is comparable to the Poisson case, i.e., very good results for $\rho > 0.5$.

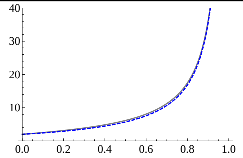
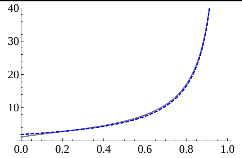
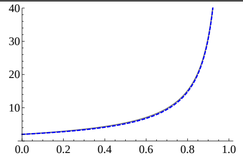
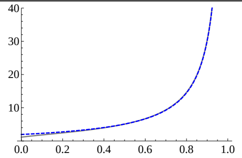
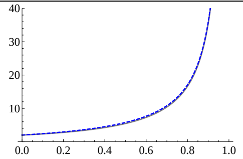
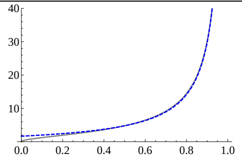
ρ	0.01	0.1	0.3	0.5	0.7	0.9	0.99	mean	standard deviation
$\mathbb{E}[W_1]$	2.0	2.5	3.9	6.2	11.2	36.1	370.4		
$\mathbb{E}[W_1^{approx}]$	2.0	2.4	3.6	5.7	10.7	35.4	369.6		
$\mathrm{sd}[W_1]$	1.3	2.0	3.6	5.9	10.9	35.3	362.7		
$\mathrm{sd}[W_1^{approx}]$	2.0	2.4	3.5	5.6	10.4	34.7	362.1		
$\mathbb{E}[W_2]$	2.0	2.4	3.5	5.4	9.8	31.2	319.1		
$\mathbb{E}[W_2^{approx}]$	2.0	2.4	3.4	5.2	9.5	30.8	318.7		
$\mathrm{sd}[W_2]$	1.2	1.8	3.1	5.1	9.4	30.3	312.4		
$\mathrm{sd}[W_2^{approx}]$	2.0	2.3	3.3	5.1	9.3	30.2	312.2		
$\mathbb{E}[W_3]$	2.0	2.3	3.4	5.5	10.4	35.5	374.8		
$\mathbb{E}[W_3^{approx}]$	2.0	2.4	3.6	5.8	10.8	35.9	375.3		
$\mathrm{sd}[W_3]$	0.4	1.2	2.7	4.8	9.2	30.2	311.6		
$\mathrm{sd}[W_3^{approx}]$	1.7	2.0	3.0	4.8	9.0	29.8	311.1		

Table 1: Results for the first numerical example. The solid grey lines in the figures correspond to the exact values, the dashed lines are approximations.

Example 2: a two-stage queueing model with customer feedback. This second example is introduced by Takács [31], 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 [31] 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. A schematic representation of this model is depicted in Figure 3.

We use the same input parameters as Takács [31]: $q = 2/3$ 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. The mean waiting times in stations one and two (corresponding to the waiting room and the service

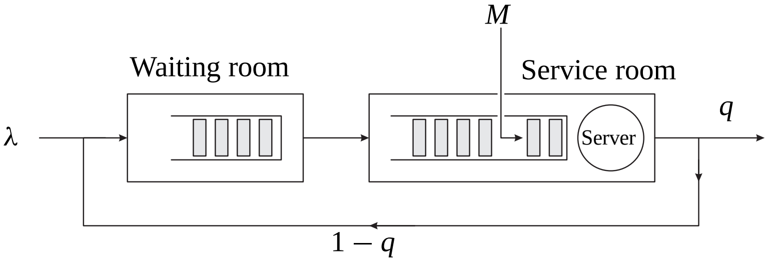


Figure 3: The two-stage queueing model with customer feedback, as discussed in Example 2.

room) are respectively

$$\mathbb{E}[W_1] = \frac{1+M}{2\mu}, \qquad \mathbb{E}[W_2] = \frac{1+7M}{6\mu}.$$

For this simple model our approximation for the mean waiting times (5.2) yields exact results.

The main purpose of this example is to illustrate how we can model a seemingly different queueing system as a special case of our model. The results are slightly different from those presented in [31], 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.

Acknowledgements

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

Appendix

A Exhaustive service

Sidi et al. [30] analysed systems with exhaustive service. They assumed last-come-first-served service, since this simplified the analysis considerably without affecting the queue length distributions. We can use the same idea, which includes using extended service times and modified transition probabilities, to compute the *cycle time distribution*. However, the first-come-first-served assumption cannot be relaxed when computing *waiting time distributions*. In this appendix we illustrate how to analyse systems with exhaustive service, while allowing some of the queues to have gated service as well. The analysis in this appendix does not reveal any new insights and is only given for completeness. We restrict ourselves to presenting the results, but we omit all proofs as they can be produced similar to the proofs in Sections 3 and 4.

In this section we use the index $e \in \{1, \dots, N\}$ to refer to an arbitrary queue with exhaustive service. The main difference between gated and exhaustive service is that 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, Sidi et al. 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). They observe that 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

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

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:

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

A.1 Queue lengths

At visit beginnings and completions. The laws of motion (3.1)-(3.2) 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 $P_i^{\mathrm{exh}}(\mathbf{z})$:

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

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

$$\begin{aligned}\widetilde{LC}^{(V_e)}(\mathbf{z}) &= \widetilde{LB}^{(V_e)}\Big(z_1, \dots, z_{e-1}, \widetilde{BP}_e\big(P_e^{\mathrm{exh}}(\mathbf{z}), \Sigma_e(\mathbf{z})\big), z_{e+1}, \dots, z_N, 1\Big), \\ \widetilde{LB}^{(R_e)}(\mathbf{z}) &= \widetilde{LC}^{(V_e)}(\mathbf{z}),\end{aligned}$$

for any exhaustively served Q_e.

At service beginnings and completions. Eisenberg's relation (3.3) remains valid for queues with exhaustive service. Note that $P_e(\mathbf{z})$ should *not* be replaced by $P_e^{\mathrm{exh}}(\mathbf{z})$ for exhaustive queues in (3.3)! Relation (3.4) should be slightly changed for queues with exhaustive service, since customers are not placed behind a gate:

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

At arbitrary moments. Equation (3.5) 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:

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

A.2 Cycle times

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 $\widetilde{B}^*_{k,i}(\omega)$.

$$\widetilde{B}^*_{k,i}(\omega) = \widetilde{BP}_{i-k}\Big(\widetilde{P}^*_{k,i}(\omega), \omega + \sum_{j=0}^{k-1}\lambda_{i-j}(1-\widetilde{B}^*_{j,i}(\omega))\Big), \qquad k=0,\dots,N; i=1,\dots,N, \tag{A.1}$$

where

$$\widetilde{P}^*_{k,i}(\omega) = 1 - \sum_{j=0}^{k-1}\frac{p_{i-k,i-j}}{1-p_{i-k,i-k}}\big(1-\widetilde{B}^*_{j,i}(\omega)\big), \qquad k=0,\dots,N; i=1,\dots,N. \tag{A.2}$$

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.

A.3 Waiting times

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

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

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

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

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 $\mathbf{B}_{\boldsymbol{0},\boldsymbol{i}}$ 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:

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

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

For each Q_i that receives gated service, we can still use (3.14)-(3.16) 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 (3.18) is determined by conditioning on the event that an arrival in Q_i takes place during $V_{i-1}, \dots, V_{i-N}$ or during $R_{i-1}, \dots, R_{i-N}$. 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 (3.24) should now become

$$\begin{aligned}\widetilde{W}_i^{(V_{i-k})}(\omega) = \widetilde{B}_{i-k}^{PR}\Big(&\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)+\lambda_i\big(1-\widetilde{B}_i(\omega)\big)+\lambda_{i-k}\big(1-\widetilde{B}^*_{k-1,i-1}(\omega)\big),\\ &\omega+\sum_{j=1}^{k-1}\lambda_{i-j}\big(1-\widetilde{B}^*_{j-1,i-1}(\omega)\big)+\lambda_{i-k}\big(1-\widetilde{B}^*_{k-1,i-1}(\omega)\big)\Big)\\ &\times\widetilde{LB}^{(B_{i-k})}\big(\mathbf{B}_{\boldsymbol{0},i}\bigotimes_{j=0}^{k-1}\mathbf{B}_{\boldsymbol{j,i-1}}\big)\prod_{j=0}^{k-1}\widetilde{R}^*_{j,i-1}(\omega)\times\frac{1-\sum_{j=0}^{k-1}p_{i-k,i-j-1}\big(1-\widetilde{B}^*_{j,i-1}(\omega)\big)}{\widetilde{B}^*_{k-1,i-1}(\omega)},\quad\text{(A.3)}\end{aligned}$$

if Q_{i-k} receives exhaustive service (and Q_i receives gated service). Compared to (3.24) 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

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

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 (3.20), should be slightly modified if Q_e receives exhaustive service. However, since the only required modification is that $\mathbf{B}_{\boldsymbol{0},i}$ should be replaced by $\mathbf{B}'_{\boldsymbol{0},i}$, 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 (3.24) if Q_{e-k} receives gated service, or (A.3) if Q_{e-k} receives exhaustive service. The only change is, once again, that $\mathbf{B}_{\boldsymbol{0},i}$ should be replaced by $\mathbf{B}'_{\boldsymbol{0},i}$. 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:

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

This concludes the analysis of a mixed model with gated and exhaustive service for $\rho < 1$. We now discuss the heavy-traffic analysis for this model.

Heavy traffic Changing the HT analysis requires substantially less work. The fluid trajectory, as depicted in Figure 1, is the same as for the gated case, except that it needs to be moved downwards such that the queue is empty at the end of the visit period. More precisely, if Q_e receives exhaustive service, then we define

$$\delta_e = \delta_e^{gated} - \hat{\rho}_e \hat{\gamma}_{e,e} \beta_e,$$

where δ_e^{gated} is the value of δ_e given by (4.1) for the case that Q_e would have received gated service. The waiting time distribution of an arbitrary fluid particle is determined similarly to the gated case, except for fluid particles arriving in Q_e during V_e, obviously. This results in the following conjecture.

Conjecture A.1 Let Q_e be an arbitrary queue with exhaustive service in the model under consideration. As $\rho \uparrow 1$, the scaled waiting time $(1-\rho)W_e$ converges in distribution to the product of a random variable having the same distribution as W_e^{fluid} and a random variable Γ having the same distribution as the limiting distribution of the scaled length-biased cycle time, $(1-\rho)\boldsymbol{C_e}$. For $\rho \uparrow 1$,

$$(1-\rho)W_e \xrightarrow{d} \begin{cases} \Gamma \times \Big((1-U_k)\hat{\rho}_k + \sum_{j=k+1}^{e-1} \hat{\rho}_j + \sum_{j=e-N+1}^{k-1} \hat{\rho}_j \hat{\gamma}_{e,j} b_e + U_k \hat{\rho}_k \hat{\gamma}_{e,k} b_e\Big) & \text{w.p. } \pi_{e,k} \\ & (k \neq e), \\ \Gamma \times U_e(\hat{\gamma}_e - \hat{\gamma}_{e,e}\hat{\rho}_e) b_e & \text{w.p. } \pi_{e,e}, \end{cases}$$

where Γ is a random variable having a Gamma distribution with parameters α and $\delta\mu$ (as defined in Section 4), and $U_1, \ldots, U_N$ are independent uniform [0, 1] distributed random variables. Note that the HT limit of W_e does not depend on whether the service disciplines in the other queues are gated or exhaustive.

Finally, we note that the LT limit remains unchanged. Since we do not consider $\mathcal{O}(\rho)$ terms, the system is always empty when a customer arrives, implying that the LT limit of the mean waiting time is independent of the service discipline.

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] M. A. A. Boon, E. M. M. Winands, I. J. B. F. Adan, and A. C. C. van Wijk. Closed-form waiting time approximations for polling systems. *Performance Evaluation*, 68:290–306, 2011.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] J. L. Dorsman, R. D. van der Mei, and E. M. M. Winands. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *Stochastic Models*, 27(2):318–332, 2011.

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

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

[18] 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.

[19] 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.

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

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

[22] 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.

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

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

[25] 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.

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

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

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

[29] 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.

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

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

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

[33] 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.

[34] 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.

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

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

[37] R. D. van der Mei and E. M. M. Winands. A note on polling models with renewal arrivals and nonzero switch-over times. *Operations Research Letters*, 36:500–505, 2008.

[38] E. M. M. Winands. On polling systems with large setups. *Operations Research Letters*, 35: 584–590, 2007.

[39] E. M. M. Winands. Branching-type polling systems with large setups. *OR Spectrum*, 33(1): 77–97, 2011.