# A Transient Analysis of Polling Systems operating under Exponential Time-Limited Service Disciplines

Roland de Haan, Ahmad Al-Hanbali, Richard J. Boucherie, Jan-Kees van Ommeren
University of Twente, Enschede, The Netherlands

March 5, 2009

**Abstract**

In the present article, we analyze a class of time-limited polling systems. In particular, we will derive a direct relation for the evolution of the joint queue-length during the course of a server visit. This will be done both for the pure and the exhaustive exponential time-limited discipline for general service time requirements and preemptive service. More specifically, service of individual customers is according to the preemptive-repeat-random strategy, i.e., if a service is interrupted, then at the next server visit a new service time will be drawn from the original service-time distribution. Moreover, we incorporate customer routing in our analysis, such that it may be applied to a large variety of queueing networks with a single server operating under one of the before-mentioned time-limited service disciplines. We study the time-limited disciplines by performing a transient analysis for the queue length at the served queue. The analysis of the pure time-limited discipline builds on several known results for the transient analysis of the M/G/1 queue. Besides, for the analysis of the exhaustive discipline, we will derive several new results for the transient analysis of an M/G/1 during a busy period. The final expressions (both for the exhaustive and pure case) that we obtain for the key relations generalize previous results by incorporating customer routing or by relaxing the exponentiality assumption on the service times. Finally, based on the interpretation of these key relations, we formulate a conjecture for the key relation for any branching-type service discipline operating under an exponential time-limit.

## 1 Introduction

Polling systems are queueing systems consisting of multiple queues served by a single server. Typically, the server visits a queue, offers service to (a part of) the customers present at this queue, and then moves to a next queue. The specific details of the system may lead to quite distinct polling models. Polling models are typically characterized by: (i) the arrival process of the customers to the system (Poisson or more general), (ii) the service requirements of the customers, (iii) the servicing policy of the server (exhaustive, gated, time-limited, etc.), (iv) the visit order of the server, and (v) the switch-over times of the server between visits to the queues. Applications of polling models are ubiquitous. For instance, traffic light systems, multiple-access protocols for communication networks (e.g., IEEE 802.11) and product-assembly systems can be modelled as a polling model. A more recent application of polling systems (see [1, 2]) is the area of wireless communication systems with mobile stations. The autonomous movements of such stations, hereby dynamically changing the network, create a specific need for studying time-limited polling models. Also due to the mobility, data transmissions may be preempted and will need to be repeated once connections are re-established. Excellent surveys on a broad class of polling models and their analysis can be found in, e.g., [3, 4, 5].

A celebrated approach to analyze polling systems is based on the construction of Markov chains at specific embedded epochs and subsequently relating the state space at these epochs (see [6]). The key relation within this approach relates the joint queue length at the end of a server visit to queue $i$ to the joint queue length at the start of the visit to queue $i$ and can be written as follows:

$$\beta^i(\mathbf{z}) = f(\alpha_i(\cdot))(\mathbf{z}) \ , \tag{1}$$

where $\beta^i(\mathbf{z})$ is the probability generating function (p.g.f.) of the joint queue length at the end of a server visit to queue $i$, $\alpha^i(\mathbf{z})$ is the p.g.f. of the joint queue length at the start of a server visit to queue $i$ and $f(\cdot)$ is a function representing the mapping between these epochs and depends on the assumed service discipline.

In the analysis of polling systems a fundamental part is played by the so-called branching property (see [7]). Polling systems which operate under service disciplines satisfying this branching property (e.g., the exhaustive and gated disciplines) are amenable to a tractable analysis, while the analysis of other disciplines (e.g., the time-limited discipline) is usually restricted to special cases or numerical approaches. This dichotomy is reflected in the function $f(\cdot)$ which for service disciplines satisfying the branching property is of a simple form, so that one obtains the following relation:

$$\beta^i(\mathbf{z}) = \alpha_i(z_1, \ldots, z_{i-1}, h_i(\mathbf{z}), z_{i-1}, \ldots, z_M) \ , \tag{2}$$

where $h_i(\mathbf{z})$ is the p.g.f. of the random population which replaces a customer served at $Q_i$ and depends on the specific service discipline. However, the time-limited discipline, according to which service is provided to customers until a time limit is reached, does not satisfy this branching property. As a result, the key relation of Eq. (1) cannot be written in the simple form of Eq. (2) and a different analytical approach is required.

Many different flavors of the time-limited discipline have been studied in the literature. The distribution of the time limit is typically assumed to be exponential but also deterministic time limits are considered. Further, the service of a customer may be either preemptive or non-preemptive. Finally, the server may depart from a queue when it becomes empty even before the time limit is reached (exhaustive service) or it may stay at the queue until the time limit is reached (non-exhaustive service). Let us next mention the literature that is closely related to our work. De Souza e Silva et al. [8] studied the key relation above for the exhaustive deterministic time-limited discipline both for preemptive and non-preemptive service. Under the assumption of exponential service times, the authors analyze the transient behavior of the system by applying uniformization techniques as to find the joint queue-length distribution $\beta^i(\mathbf{z})$. Leung [9] analyzed the key relation for the exhaustive exponential time-limited discipline and non-preemptive service. This was done in a recursive manner by conditioning on specific intermediate events during a server visit. Eliazar and Yechiali [10, 11] studied the exhaustive exponential time-limited discipline for preemptive service. Observing that upon successful service completion at a queue the busy period in fact regenerates, the authors could obtain a closed-form relation between the joint queue length at the end and start of a server visit of the following form:

$$\beta^i(\mathbf{z}) = c(\mathbf{z}) \cdot (\alpha_i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + \alpha^i(\mathbf{z}_i^*) \ , \tag{3}$$

where $\alpha^i(\mathbf{z}_i^*) := \alpha_i(z_1, \ldots, z_{i-1}, k_i(\mathbf{z}), z_{i-1}, \ldots, z_M)$, and $c(\mathbf{z})$ and $k_i(\mathbf{z})$ are functions of $\mathbf{z}$ with $k_i(\mathbf{z})$ being related to the length of the busy period of a customer at $Q_i$. Under the assumption of exponential service times, Al Hanbali et al. [12] derived a similar relation between $\beta^i(\mathbf{z})$ and $\alpha^i(\mathbf{z})$ for the non-exhaustive exponential time-limited discipline using a matrix geometric approach. Along the same approach, the authors also rederived Eq. (3) for the exhaustive discipline under exponential service times.

Complementary to the key relation, Eq. (1), there exists a relation between $\beta^i(\mathbf{z})$ and $\alpha^j(\mathbf{z})$ which couples the queue length at the start of a visit to queue $j$ to the queue length at the end of a visit to queue $i$:

$$\alpha^j(\mathbf{z}) = C_{ij}(\mathbf{z}) \cdot \beta^i(\mathbf{z}), \; j \neq i \; , \tag{4}$$

where $C_{ij}(\mathbf{z})$ denotes the p.g.f. of the number of arrivals to all queues during the switch-over time of the server from queue $i$ to queue $j$. Clearly, Eq. (4) is independent of the service discipline.

The relations of Eq. (1) and (4) for all queues in the system together give rise to a system of equations which may be solved in an iterative fashion. For disciplines satisfying the branching property, this leads to a closed-form solution for the joint queue-length distributions at the embedded epochs. Although also Eq. (3) may seem quite explicit, the system of equations obtained for other service disciplines does typically not lead to closed-form expressions for the queue-length distribution, so that one must resort to numerical solution methods.

In this work, we will study the key relation between $\beta^i(\mathbf{z})$ and $\alpha^i(\mathbf{z})$, i.e., we analyze how the joint queue-length evolves during the course of a server visit. This will be done both for the exhaustive and the non-exhaustive exponential time-limited discipline for general service time requirements. Service of individual customers is according to the *preemptive-repeat-random* strategy, i.e., if a service is interrupted, then at the next server visit a new service time will be drawn from the original service-time distribution. The motivation for this latter choice is that the transmission time of data in a wireless environment is highly related to the randomness of the communication medium and that the size of the data plays only a minor role. Hence, in light of this application that we have in mind, it is more appropriate to redraw a new random service time rather than to retain the original service time upon a service interruption. Moreover, we incorporate customer routing in our analysis, such that it may be applied to any kind of queueing network with a single server operating under one of the before-mentioned time-limited service disciplines. The analysis of the non-exhaustive discipline builds on several known results for the transient analysis of the M/G/1 queue. On the contrary, to analyze the exhaustive discipline, we will derive several new results for the transient analysis of an M/G/1 during a busy period. The final expressions (both for the exhaustive and non-exhaustive case) that we obtain for the key relations are of the form:

$$\beta^i(\mathbf{z}) = d_1(\mathbf{z}) \cdot (\alpha_i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*) \; , \tag{5}$$

where $\alpha^i(\mathbf{z}_i^*) := \alpha_i(z_1, \ldots, z_{i-1}, l_i(\mathbf{z}), z_{i-1}, \ldots, z_M)$, $d_1(\mathbf{z})$ and $d_2(\mathbf{z})$ are functions which are largely determined by the Laplace-Stieltjes Transform (LST) of the service-time distribution, and $l_i(\mathbf{z})$ is related to the length of the busy period of a customer at $Q_i$. These relations generalize previous results by incorporating customer routing ([10] and [12]) and by relaxing the exponentiality assumption on the service times [12].

The rest of this work is organized as follows. We describe the model and the notation in Sect. 2. The key relations for the non-exhaustive and the exhaustive exponential time-limited discipline are presented in Sects. 3 and 4, respectively. In Sect. A, we study the transient behavior for a M/G/1 queue during a busy period. We conclude this work with a discussion on the final results for the key relations in Sect. 5. The complete proofs of the key relations are given Appendices B and C.

## 2 Model and notation

Consider a system of $M$ queues denoted by $Q_1, \dots, Q_M$, which are served by a single server at unit rate. Customers arrive to $Q_i$ according to a Poisson arrival stream with rate $\lambda_i,\ i = 1, \dots, M$. The service requirements $S_i$ of a customer at $Q_i$ are generally distributed with mean $b_i$. The switch-over times for the server to move from $Q_i$ to $Q_j,\ i, j = 1, \dots, M$, are denoted by $c_{ij}$.

Customers are served at the queue according to a specific service discipline. In this work, we will focus on two service disciplines, viz.,

- (i) non-exhaustive exponential time-limited discipline ;
- (ii) exhaustive exponential time-limited discipline .

According to both disciplines, the server will at most visit a queue for an exponential amount of time $T_i$ which is exponentially distributed with rate $\xi_i$. However, under the exhaustive discipline the server will move to a next queue as soon as the queue becomes empty, while under the non-exhaustive discipline the server remains at the queue until the timer expires. If the server is still present upon expiration of the timer, it moves immediately to a next queue without completing any on-going service. At the next visit of the server to the queue, a new service time will be drawn for an interrupted customer from the original service-time distribution, i.e., service occurs according to the so-called *preemptive-repeat-random* strategy.

Customers who have completed their service at $Q_i,\ i = 1, \dots, M$, will join $Q_j,\ j = 1, \dots, M$ with probability $r_{ij} \geq 0$ and with probability $r_{i0} \geq 0$ they will leave the system. Clearly, these routing probabilities $r_{ij}$ must satisfy $\sum_{j=0}^{M} r_{ij} = 1,\ i = 1, \dots, M$. We will assume in the sequel that $r_{ii} = 0$, i.e., no self loops are allowed. However, we note that $r_{ii} > 0$ might also be incorporated in the model (see Remark 1). Finally, we let $r^i(\mathbf{z})$ denote the p.g.f. of the number of arrivals to all queues generated by a single departing customer at $Q_i$, i.e., $r^i(\mathbf{z}) = r_{i0} + \sum_j r_{ij} z_j$.

The server serves the queues according to the periodic polling strategy. Without loss of generality (w.l.o.g.) we define a cycle as the time period between two consecutive polling instants at the 1st stage (or visit) of the cycle. A cycle consists of $a$ stages and we denote by $t(j),\ j = 1, \dots, a$, the queue served during stage $j$ of the cycle. Further, the number of times $Q_i$ is visited during a cycle is denoted by $a_i,\ i = 1, \dots, M$, with $a_i \geq 1$ and $\sum_{i=1}^{M} a_i = a$.

**Remark 1.** *The case $r_{ii} > 0$ may be incorporated in the model by appropriately scaling the service rates and the routing probabilities at a queue. To be precise, the service time should be scaled such that its mean, denoted by $b'_i$, equals $b_i/(1 - r_{ii})$. The scaled routing probabilities, $r'_{ij}$, should be set to $r_{ij}/(1 - r_{ii}),\ j \neq i$, while $r'_{ii}$ should be set to zero. In this modified system, the server serves each arriving customer only once, but as each brings more work to the queue the total effective amount of work arriving per time unit to the queue remains the same as for the original system. Finally, using a sample-path comparison, it can readily be seen that the queue-length distribution of the modified system is equal to the one of the original system.*

Below we introduce the notation that will be used throughout.

- $x_t$; number of customers at time $t$ at $Q_i$;
- $z_n$; number of customers left behind by the $n$th departing customer from $Q_i$;
- $r'_n$; time of the $n$th departure from $Q_i$;
- $D(t)$; number of departures from $Q_i$ in $[0, t)$;
- $A_i(t)$; number of arrivals to $Q_i$ in $[0, t)$;
- $I_i$; exponentially distributed random variable with parameter $\lambda_i$ denoting the interarrival time to $Q_i$;
- $S_i$; generally distributed random variable denoting the service time at $Q_i$;
- $\mathbf{1}_{\{A\}}$; indicator function of event $A$;
- $\tilde{X}(\cdot)$; LST of random variable $X$;
- $\mu(s, y)$; root $x$ with the smallest absolute value less than one of $x = y \cdot \tilde{S}_i(s + \lambda_i(1 - x))$;
- $\mathbf{N}_i^s$; number of customers at all queues at the start of a server visit to $Q_i$;
- $\mathbf{N}_i^e$; number of customers at all queues at the end of a server visit to $Q_i$;
- $N_{i,j}(t)$; number of customers at $Q_j$ at time $t$ during a server visit to $Q_i$;
- $\alpha^i(\mathbf{z})$; p.g.f. of $\mathbf{N}_i^s$;
- $\beta^i(\mathbf{z})$; p.g.f. of $\mathbf{N}_i^e$.

# 3 Analysis of the non-exhaustive time-limited service discipline

In this section, we analyze the non-exhaustive time-limited discipline. Under this discipline, the server will only depart from the queue when the time limit has been reached. It should be stressed that the server will not leave the queue when it becomes empty. We will derive an expression for $\beta^i(\mathbf{z})$, the p.g.f. for the number of customers at all queues at the instant that the server leaves $Q_i$, in terms of the number present at the start of the visit, $\alpha^i(\mathbf{z})$. Here, we present only the essential analytical steps and the main result. The proofs will be given in Appendix B.

A necessary and sufficient condition for the stability of a polling system with the server operating under the pure exponential time-limited discipline is given in the following theorem.

**Theorem 1** (Pure exponential time-limited discipline)**.**

$$\textit{System is stable} \iff \rho_i < \kappa_i, \ \forall_{i \in \{1,\ldots,M\}} \ , \tag{6}$$

*where*

$$\rho_i = \lambda_i \cdot \frac{1 - \tilde{S}_i(\xi_i)}{\xi_i \cdot \tilde{S}_i(\xi_i)} \ , \tag{7}$$

$$\kappa_i = \frac{a_i/\xi_i}{\sum_{j=1}^{M} a_j/\xi_j + \sum_{k=1}^{a} c_{t(k),t(k+1)}} \ . \tag{8}$$

*Proof.* It is well-known that for a single queue the nonsaturation condition is both a necessary and sufficient condition for stability, i.e.,

$$Q_i \text{ is stable } \iff \rho_i < \kappa_i,\ i = 1, \dots, M\ , \tag{9}$$

where $\rho_i$ is the mean effective amount of work arriving per time unit to $Q_i$ and $\kappa_i$ is the availability fraction of the server at $Q_i$.

Consider first the mean effective amount of work arriving per time unit to $Q_i$. This amount is determined by the total number of customers arriving per time unit $\lambda_i$ and the mean effective amount of work each individual brings for the server $\mathbb{E}[S_E]$ as follows:

$$\rho_i = \gamma_i \cdot \mathbb{E}[S_E]\ . \tag{10}$$

The quantity $\mathbb{E}[S_E]$ is in fact the mean total time the server spends on serving a customer at $Q_i$ including any interrupted services. Noting that the number of interruptions per customer is geometrically distributed, it can be found via simple calculus that:

$$\mathbb{E}[S_E] = \frac{1 - \tilde{S}_i(\xi_i)}{\xi_i \cdot \tilde{S}_i(\xi_i)}\ . \tag{11}$$

The availability fraction of the server $\kappa_i$ is fully specified by the mean visit times, the visit frequencies and the switch-over times between the queues. Notice that a complete cycle consists of $a_i$ visits to $Q_i,\ i = 1, \dots, M$, and the switch-over times between the queues. It then readily follows for the availability fraction of the server at $Q_i$:

$$\kappa_i = \frac{a_i/\xi_i}{\sum_{j=1}^{M} a_j/\xi_j + \sum_{k=1}^{a} c_{t(k),t(k+1)}}\ . \tag{12}$$

It is good to notice that the fraction $\kappa_i$ is independent of the load at the queues. The observation that the system is stable if and only if all the queues in the system are stable completes the proof. □

### 3.1 Relating $\beta^i(\mathbf{z})$ to $\alpha^i(\mathbf{z})$

Consider a visit of the server to $Q_i$. During such a visit, the queue-length process at $Q_i$ is a birth-and-death process, while the queue-length process at the other queues is a pure birth-process. Notice that arrivals to $Q_j,\ j \neq i$, may be both exogenous and endogenous (from $Q_i$). Our interest is in the number of customers at time $t$ given a certain initial number of customers at $Q_i$. Moreover, to include customer routing in the analysis, we need to keep track of the number of departures during a visit. Notice that to record this number of departures, it is not sufficient to know the number of customers at $Q_i$ at the beginning and the end of a visit. Therefore, we will focus on the transient probabilities $p_{hk}^{(n)}(t)$ which are defined as follows:

$$p_{hk}^{(n)}(t) := \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n | x_0 = h), & h, k, n = 0, 1, \dots, \\ 0, & \text{otherwise}\ . \end{cases}$$

where for notational convenience the dependence of $p_{hk}^{(n)}(t)$ on $Q_i$ is suppressed. We will relate these probabilities to the transient probabilities for the standard M/G/1 queue, which we denote by $P_{hj}^{(n)}(t)$. These time-dependent conditional probabilities which incorporate also the number of departures until time $t$ are defined for $n = 1, 2, \ldots,\ h, j = 0, 1, \ldots,$ and $t > 0$ as [13]:

$$P_{hj}^{(n)}(t) := \mathbb{P}(z_n = j,\ r'_n \leq t | z_0 = h)\ , \tag{13}$$

where it is assumed that at time $t = 0$ the 0-th customer left the queue. We consider the function $\pi_h(r, s, y)$ which is defined in terms of $P_{hj}^{(n)}(t)$ as follows:

$$\pi_h(r, s, y) := \sum_{n=1}^{\infty} y^n \sum_{j=0}^{\infty} r^j \int_0^{\infty} e^{-st} dP_{hj}^{(n)}(t),\ h = 0, 1, \ldots\ , \tag{14}$$

and which is explicitly provided in Cohen [13] as

$$\pi_h(r, s, y) = \frac{y \cdot \tilde{S}_i(s + \lambda_i(1 - r))}{r - y \cdot \tilde{S}_i(s + \lambda_i(1 - r))} \cdot \left\{ r^h - \frac{\lambda_i(1 - r) + s}{\lambda_i(1 - \mu(s, y)) + s} \cdot \mu^h(s, y) \right\},\ h = 0, 1, \ldots\ , \tag{15}$$

where $\mu(s, y)$ is the root $x$ with the smallest absolute value less than one of $x = y \cdot \tilde{S}_i(s + \lambda_i(1 - x))$. Notice that $\mu(s, 1)$ equals the LST with parameter $s$ of the length of the busy period at $Q_i$.

To take advantage of this explicit result, we will first present an explicit expression for the transient probabilities $p_{hk}^{(n)}(t)$ in terms of $P_{hj}^{(n)}(t)$. For convenience, we define:

$$\begin{aligned} F_k^{(0)}(t) &= \mathbf{1}_{\{k=0\}} \mathbb{P}(A_i(t) = 0,\ I_i > t) + \mathbf{1}_{\{k \geq 1\}} \mathbb{P}(A_i(t) = k,\ I_i + S_i > t),\ k = 0, 1, \ldots\ , & (16) \\ F_k^{(j)}(t) &= \mathbb{P}(A_i(t) = k - j,\ S_i > t),\ j = 1, 2, \ldots,\ k = j, j + 1, \ldots\ . & (17) \end{aligned}$$

That is, $F_k^{(j)}(t)$ refers to $k - j$ exogenous arrivals to $Q_i$ during a server visit to $Q_i$ initiated with $j$ customers and which duration is shorter than a service time $S_i$ meaning that a service is interrupted (except when $j = k = 0$). In the special case $j = 0$, we need to account for the fact that first an arrival should occur before any service may start at all. Then, we can relate $p_{hk}^{(n)}(t)$ to $P_{hj}^{(n)}(t)$ for $n = 1, 2, \ldots,\ h, k = 0, 1, \ldots,$ and $t > 0$ as follows:

**Lemma 1.**

$$p_{hk}^{(n)}(t) = \int_{u=0}^{t} F_k^{(0)}(t - u) dP_{h0}^{(n)}(u) + \sum_{j=1}^{k} \int_{u=0}^{t} F_k^{(j)}(t - u) dP_{hj}^{(n)}(u)\ . \tag{18}$$

To retrieve the terms $\pi_h(r, s, y)$, we take the LST of $p_{hk}^{(n)}(t)$ (see Remark 3). Next, we will take the generating function of this expression with respect to the number of customers at the end of a server visit. Notice that our interest here is specifically in this number rather than in the number at the time of the $n$th departure, since the server only leaves upon expiration of the timer. In a final step, we take the generating function with respect to the number of departures until time $t$ as to obtain an expression for $p_{hk}^{(n)}(t)$ in terms of $\pi_h(r, s, y)$. These consecutive steps provide us with the following result.

**Lemma 2.**

$$\begin{aligned}
&\sum_{n=1}^{\infty} y^n \sum_{k=0}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dp_{hk}^{(n)}(t) && (19)\\
= \quad & \frac{s}{\lambda_i(1-r)+s} \cdot \frac{\lambda_i(1-r\cdot\tilde{S}_i(\lambda_i(1-r)+s))+s}{\lambda_i+s} \cdot \pi_{h0}(s,y)\\
+ \quad & \frac{s}{\lambda_i(1-r)+s} \cdot (1-\tilde{S}_i(\lambda_i(1-r)+s)) \cdot (\pi_h(r,s,y)-\pi_{h0}(s,y)) \, , \;\; h=0,1,\dots \; ,
\end{aligned}$$

*where the terms $\pi_{h0}(s,y)$ are given by (see [13]),*

$$\pi_{00}(s,y) = \frac{\lambda_i}{\lambda_i(1-\mu(s,y))+s} \cdot \mu(s,y) \; , \tag{20}$$

$$\pi_{h0}(s,y) = \frac{\lambda_i+s}{\lambda_i(1-\mu(s,y))+s} \cdot \mu^h(s,y), \;\; h=1,2,\dots \; . \tag{21}$$

The right-hand side of Eq. (19) can be interpreted as follows. The first part refers to the case that upon the $n$th departure zero customers are left behind, while the second part refers to a strictly positive number left behind by the $n$th departing customer. Moreover, the second part can be decomposed in two independent components: $\pi_h(r,s,y)-\pi_{h0}(s,y)$ accounts for the queue-length evolution until $n$th departure and the other component for the queue-length evolution during the final, interrupted service. A similar reasoning holds for the first part.

Thus, we have related the transient probabilities of our interest to known results for the M/G/1 queue. Next, by unconditioning on the system state at the start of a visit and incorporating the expressions above into the definition of $\beta^i(\mathbf{z})$, we obtain the main result of this section for the p.g.f. of the joint queue-length at the end of a server visit under the non-exhaustive exponential time-limited discipline.

**Theorem 2.**

$$\beta^i(\boldsymbol{z}) = d_1^{NE}(\boldsymbol{z}) \cdot (\alpha_i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + d_2^{NE}(\boldsymbol{z}) \cdot \alpha^i(\boldsymbol{z}_i^*) \; , \tag{22}$$

*where*

$$d_1^{NE}(\boldsymbol{z}) = \frac{\xi_i}{z_i - r^i(\boldsymbol{z}) \cdot \tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)} \cdot \frac{z_i \cdot (1-\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*))}{\lambda_i(1-z_i)+\xi_i^*} \; , \tag{23}$$

$$d_2^{NE}(\boldsymbol{z}) = \frac{\xi_i}{z_i - r^i(\boldsymbol{z}) \cdot \tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)} \cdot \frac{(z_i - r^i(\boldsymbol{z})) \cdot \tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)}{\lambda_i(1-\mu(\xi_i^*, r^i(\boldsymbol{z})))+\xi_i^*} + d_1^{NE}(\boldsymbol{z}) \; , \tag{24}$$

$$\xi_i^* = \xi_i + \sum_{j\neq i} \lambda_j(1-z_j) \; , \tag{25}$$

*and* $\alpha^i(\boldsymbol{z}_i^*) := \alpha^i(z_1, \dots, z_{i-1}, \mu_i(\xi_i^*, r^i(\boldsymbol{z})), z_{i+1}, \dots, z_M)$.

**Remark 2** (Exponential service times)**.** *For the case of exponential service times at $Q_i$ (with rate $1/b_i$), it can be shown that Eq. (43) can be rewritten to:*

$$\beta^i(\boldsymbol{z}) = \frac{\xi_i \cdot z_i}{V_i(\boldsymbol{z})} \cdot (\alpha_i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + \frac{\xi_i \cdot r^i(\boldsymbol{z}) \cdot (\mu(\xi_i^*, r^i(\boldsymbol{z})) - z_i)}{V_i(\boldsymbol{z}) \cdot (\mu(\xi_i^*, r^i(\boldsymbol{z})) - r^i(\boldsymbol{z}))} \cdot \alpha^i(\boldsymbol{z}_i^*) \; , \tag{26}$$

*where*

$$V_i(\boldsymbol{z}) = -\lambda_i z_i^2 + (1/b_i + \lambda_i + \sum_{j \neq i} \lambda_j (1 - z_j) + \xi_i) z_i - r^i(\boldsymbol{z})/b_i \ . \tag{27}$$

*We note that Eq. (26) generalizes the result for the special case $r^i(\boldsymbol{z}) = 1$ (i.e., no customer routing) given in [12].*

**Remark 3** (Exponential time limit)**.** *The step of taking the LST of $p_{hk}^{(n)}(t)$ corresponds to unconditioning over the exponentially distributed visit time. This shows that the assumption on the visit time plays a crucial role in the analysis.*

# 4 Analysis of the exhaustive time-limited service discipline

Let us next consider the exhaustive time-limited discipline. Notice that under this discipline the server will depart from the queue when it becomes empty or when the time limit has been reached, whichever occurs first. Again, we will derive an expression for $\beta^i(\mathbf{z})$, the p.g.f. for the number of customers at all queues at the instant that the server leaves $Q_i$. This will be done in terms of the number present at the start of the visit, $\alpha^i(\mathbf{z})$. As in the previous section, we present here only the main analytical steps and the final result. The proofs will be given in Appendix C.

## 4.1 Stability

The polling system is stable if there exists a stationary regime in which each customer in the system can be served in a finite period of time. For the exhaustive time-limited discipline, service capacity can be exchanged between the queues. This suggests that stability can be considered for the system as a whole. However, as the visit time to each queue is bounded by the timer, the occupancy of individual queues also plays a role.

A necessary and sufficient condition for the stability of a polling system with the server operating under the exhaustive exponential time-limited discipline is given in the following theorem.

**Theorem 3** (Exhaustive exponential time-limited discipline)**.**

$$\text{System is stable} \iff \rho + \max_{1 \leq i \leq M} \left( \frac{\lambda_i}{\mathbb{E}[G_i^{*-}]} \right) \cdot c_T < 1 \ , \tag{28}$$

*where $c_T$ is the mean total switch-over time during a cycle and $\mathbb{E}[G_i^{*-}]$ denotes the mean maximum number of served customers at $Q_i$ during a cycle given by:*

$$\mathbb{E}[G_i^{*-}] = \frac{a_i \cdot \tilde{S}_i(\xi_i)}{1 - \tilde{S}_i(\xi_i)} \ , \qquad , \tag{29}$$

## 4.2 Relating $\beta^i(\mathbf{z})$ to $\alpha^i(\mathbf{z})$

Under the exhaustive time-limited discipline, the server may leave a queue for two reasons, viz., the server departs due to the queue being empty or due to the timer expiring. Let $\{empty\}$ and $\{timer\}$ denote the corresponding server events. Recall that $\mathbf{N}_i^s$ and $\mathbf{N}_i^e$ denote the multi-dimensional r.v. of the number of customers at all queues at the start and the end of a visit to $Q_i$, respectively. The p.g.f. of $\mathbf{N}_i^e$, $\beta_i(\mathbf{z})$, can be decomposed in two parts depending on the reason of a server departure as the server departs only if the queue is empty or if the timer expires. Moreover, these events are readily seen to be mutually exclusive (service-time distribution and timer distribution are both continuous distributions, so that the probability of the given events occurring simultaneously is zero). Hence, the p.g.f. for the number of customers at the end of a visit period to $Q_i$ satisfies,

$$\beta_i(\mathbf{z}) = \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}] = \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}] + \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}] \ . \tag{30}$$

Next, in the Sects. 4.2.1 and 4.2.2, we will derive the conditional p.g.f.'s $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}|\mathbf{N}_i^s = \mathbf{n}]$ and $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}|\mathbf{N}_i^s = \mathbf{n}]$, where $\mathbf{n}$ denotes the vector $(n_1, \dots, n_M)$. Finally, we will uncondition the expressions to get our main result in Sect. 4.2.3.

### 4.2.1 $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}|\mathbf{N}_i^s = \mathbf{n}]$

We note that in case the $\{empty\}$ event occurs the queue may be empty upon arrival of the server or become empty at a departure of a customer. If the server finds an empty queue upon arrival, then clearly $\mathbf{N}_i^e = \mathbf{N}_i^s$. Else, if the queue is nonempty, then the evolution of queue-length process during the visit is strongly related to the length of a busy period in a standard M/G/1 queue. This is formalized in the following proposition.

**Proposition 1.** *The joint conditional p.g.f. of the number of customers at the end of a visit period to $Q_i$ and the server departs due to the queue being empty satisfies,*

$$\mathbb{E}[\boldsymbol{z}^{\boldsymbol{N}_i^e}\boldsymbol{1}_{\{empty\}}|\boldsymbol{N}_i^s = \boldsymbol{n}] = \mu_i^{n_i}(\xi_i^*, r^i(\boldsymbol{z})) \cdot \prod_{j\neq i} z_j^{n_j} \ , \tag{31}$$

*where*

$$\xi_i^* = \xi_i + \sum_{j\neq i} \lambda_j(1 - z_j) \ . \tag{32}$$

### 4.2.2 $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s = \mathbf{n}]$

We note that in case the $\{timer\}$ event occurs the queue must be nonempty upon arrival of the server, then it remains nonempty during the course of the visit and it is still nonempty at the expiration of the timer. The analysis of this case builds on the work of Cohen for the transient analysis of the M/G/1 queue. However, contrary to the analysis for the non-exhaustive time-limited discipline, we cannot directly apply the formulae derived in [13]. This is due to the fact that we need specifically to account for not entering the state with zero customers at $Q_i$ during the course of a server visit. Below, we state the transient probabilities of interest and several related expressions. Next, using these expressions, we will derive $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s = \mathbf{n}]$.

We consider the conditional joint queue-length distribution at time $t > 0$ given an initial number of customers at time $t = 0$ and given that the server is at $Q_i$. It is good to notice that during a server visit to $Q_i$ the queue-length process at the other queues is simply a pure-birth process. Hence, we neglect the other queues for the moment and concentrate on the marginal queue-length probabilities for $Q_i$, denoted by $q_{hk}^{(n)}(t)$, which we define as:

$$q_{hk}^{(n)}(t) := \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n,\ x_v > 0,\ 0 < v < t | x_0 = h), & n = 0, 1, \dots,\quad h, k = 1, 2, \dots, \\ 0, & \text{otherwise} \ , \end{cases}$$

where for notational convenience the dependence of $q_{hk}^{(n)}(t)$ on $Q_i$ is suppressed. For completeness, let us recall the definition of the probabilities $P_{hj}^{(n)}(t)$, for $n = 1, 2, \ldots,\ h, j = 0, 1, \ldots$ and $t > 0$,

$$P_{hj}^{(n)}(t) := \mathbb{P}(z_n = j,\ r'_n < t | z_0 = h)\ . \tag{33}$$

Analogously, we define $R_{hj}^{(n)}(t)$, for $h, j, n = 1, 2, \ldots$, and $t > 0$,

$$R_{hj}^{(n)}(t) := \mathbb{P}(z_n = j,\ r'_n < t,\ z_k > 0,\ 0 < k < n | z_0 = h)\ , \tag{34}$$

where it is assumed that at time $t = 0$ a new service starts. We note that $R_{hj}^{(n)}(t)$ is only defined for $h, j = 1, 2, \ldots$ . This is due to the fact that the event of a server arriving to an empty queue (i.e., $h = 0$) and the event of the $n$th customer leaving an empty queue behind (i.e., $j = 0$) are never considered as $\{timer\}$ events, but always as $\{empty\}$ events.

We consider the function $\gamma_h(r, s, y)$ which is defined in terms of $R_{hj}^{(n)}(t)$ as follows:

$$\gamma_h(r, s, y) := \sum_{n=1}^{\infty} y^n \sum_{j=1}^{\infty} r^j \int_0^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h = 1, 2, \ldots\ , \tag{35}$$

and which is explicitly given (see Sect. A for the derivation) for $h = 1, 2, \ldots$, as

$$\gamma_h(r, s, y) = \frac{r}{r - y \cdot \tilde{S}_i\left(\lambda_i(1 - r) + s\right)} \cdot \left(-\mu^h(s, y) + y \cdot \tilde{S}_i\left(\lambda_i(1 - r) + s\right) \cdot r^{h-1}\right)\ . \tag{36}$$

Analogously to the approach in the previous section, we intend to utilize the explicit expressions for $\gamma_h(r, s, y)$. To this end, we will start by relating the transient probabilities $q_{hk}^{(n)}(t)$ to the time-dependent probabilities $R_{hj}^{(n)}(t)$ at embedded epochs of service completion instants. For convenience, we recall that:

$$F_k^{(j)}(t) = \mathbb{P}(A_i(t) = k - j,\ S > t),\ j = 1, 2, \ldots\ ,\ k = j, j + 1, \ldots\ , \tag{37}$$

that is, $F_k^{(j)}(t)$ refers to the number of arrivals to $Q_i$ during a visit to $Q_i$ initiated with $j$ customers and which duration is shorter than a service time $S_i$. The specific relation between $q_{hk}^{(n)}(t)$ and $R_{hj}^{(n)}(t)$ is then given in the following lemma.

**Lemma 3.**

$$q_{hk}^{(n)}(t) = \int_{u=0}^{t} \sum_{j=1}^{k} F_k^{(j)}(t - u) dR_{hj}^{(n)}(u),\ n = 1, 2, \ldots,\ h, k = 1, 2, \ldots\ . \tag{38}$$

Again, to obtain the terms $\gamma_h(r, s, y)$, we take the LST of $q_{hk}^{(n)}(t)$ (see Remark 3). Next, we take the generating function with respect to the number of customers at the end of the server visit of the resulting expression and finally we take the generating function with respect to the number of departures. Hence, we obtain the following result.

**Lemma 4.**

$$\begin{aligned}
&\sum_{n=1}^{\infty} y^n \sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(n)}(t) && (39)\\
= \;& \gamma_h(r,s,y) \cdot \frac{s}{\lambda_i(1-r)+s} \cdot (1-\tilde{S}_i(\lambda_i(1-r)+s)), \;\; h=1,2,\dots \;. && (40)
\end{aligned}$$

The right-hand side of Eq. (40) can be recognized as a convolution of two independent parts. The first part, $\gamma_h(r,s,y)$, refers to the queue length at the instant of the final (successful) service completion during the visit, while the other part refers to number of arrivals during an interrupted service.

Next, we can present the explicit expression for the joint conditional p.g.f. of the number of customers at all queues at the end of a visit when server departure is due to the timer expiration. The condition is on the number of customers present at the start of the visit.

**Proposition 2.**

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{z}^{\boldsymbol{N}_i^e} \boldsymbol{1}_{\{timer\}} | \boldsymbol{N}_i^s = \boldsymbol{n}] && (41)\\
= \;& \frac{\xi_i \cdot z_i \cdot (1-\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*))}{[\lambda_i(1-z_i)+\xi_i^*] \cdot [z_i - r^i(\boldsymbol{z}) \cdot \tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)]} \cdot \left(z_i^{n_i} - \mu^{n_i}(\xi_i^*, r^i(\boldsymbol{z}))\right) \cdot \prod_{j\neq i} z_j^{n_j} \;,
\end{aligned}$$

*where*

$$\xi_i^* = \xi_i + \sum_{j\neq i} \lambda_j(1-z_j) \;. \tag{42}$$

### 4.2.3 $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}]$

Combining the two conditional results of Eqs. (31) and (41), we obtain our main result of this section for the exhaustive exponential time-limited service discipline.

**Theorem 4.**

$$\beta^i(\boldsymbol{z}) \;=\; d_1^E(\boldsymbol{z}) \cdot (\alpha_i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + d_2^E(\boldsymbol{z}) \cdot \alpha^i(\boldsymbol{z}_i^*) \;, \tag{43}$$

*where*

$$\begin{aligned}
d_1^E(\boldsymbol{z}) &= d_1^{NE}(\boldsymbol{z}) \;, && (44)\\
d_2^E(\boldsymbol{z}) &= 1 \;, && (45)\\
\xi_i^* &= \xi_i + \sum_{j\neq i} \lambda_j(1-z_j) \;, && (46)
\end{aligned}$$

*and* $\alpha^i(\boldsymbol{z}_i^*) := \alpha^i(z_1,\dots,z_{i-1},\mu_i(\xi_i^*, r^i(\boldsymbol{z})), z_{i+1},\dots,z_M)$.

We note that Eq. (43) generalizes the result for the special case $r^i(\mathbf{z}) = 1$ (i.e., no customer routing) given in [10].

**Remark 4** (Exponential service times)**.** *For the case of exponential service times at $Q_i$ (with rate $1/b_i$), it can be shown that Eq. (43) can be rewritten to:*

$$\beta^i(\boldsymbol{z}) = \frac{\xi_i \cdot z_i}{V_i(\boldsymbol{z})} \cdot (\alpha_i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + \alpha^i(\boldsymbol{z}_i^*) \ , \tag{47}$$

*where $V_i(\boldsymbol{z})$ is given in Eq. (27). We note that thus Eq. (47) generalizes the result for the special case $r^i(\boldsymbol{z}) = 1$ (i.e., no customer routing) given in [12].*

**Remark 5** (Exhaustive service discipline)**.** *We note that in the limit case of $\xi_i \downarrow 0$ the time limit is of infinite length. Hence, in this case (assuming a stable queue), the server will always depart due to $Q_i$ being empty. It can readily be found that for $\lim \xi_i \downarrow 0$ and $r^i(\boldsymbol{z}) = 1$ the following expression for $\beta^i(\boldsymbol{z})$ is obtained:*

$$\beta^i(\boldsymbol{z}) = \alpha^i(\boldsymbol{z}_i^*) \ , \tag{48}$$

*where $\alpha^i(\boldsymbol{z}_i^*) := \alpha^i(z_1, \dots, z_{i-1}, \mu_i(\sum_{j \neq i} \lambda_j(1 - z_j), 1), z_{i+1}, \dots, z_M)$. This result matches the well-known result for the exhaustive service discipline.*

# 5 Discussion

The final results for the exhaustive exponential time-limited discipline (E-TL) and the non-exhaustive exponential time-limited discipline (NE-TL) are similar. More specifically, these results can be written in the following form:

$$\text{NE-TL: } \beta^i(\mathbf{z}) = d_1^{NE}(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2^{NE}(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*) \tag{49}$$

$$\text{E-TL: } \beta^i(\mathbf{z}) = d_1^{E}(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2^{E}(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*), \tag{50}$$

where $d_1^E(\mathbf{z})$ $(= d_1^{NE}(\mathbf{z}))$ is given in Eq. (23), $d_2^E(\mathbf{z}) = 1$ and $d_2^{NE}(\mathbf{z})$ is given by

$$\begin{aligned} d_2^{NE}(\mathbf{z}) \quad = \quad & \frac{\xi_i}{z_i - r^i(\mathbf{z}) \cdot \tilde{S}_i(\lambda_i(1 - z_i) + \xi_i^*)} \\ & \times \left\{ \frac{\tilde{S}_i(\lambda_i(1 - z_i) + \xi_i^*) \cdot (z_i - r^i(\mathbf{z}))}{\lambda_i(1 - \mu_i(\xi_i^*, r^i(\mathbf{z}))) + \xi_i^*} + \frac{z_i(1 - \tilde{S}_i(\lambda_i(1 - z_i) + \xi_i^*))}{\lambda_i(1 - z_i) + \xi_i^*} \right\} \ . \end{aligned} \tag{51}$$

Equations (49) and (50) can be interpreted as follows. Consider a visit of the server to $Q_i$. Regarding the timer, it may occur that (i) the timer expires before $Q_i$ gets empty for the first time, or (ii) the timer expires only after $Q_i$ becomes empty for the first time. It is readily seen that the queue-length process is identical for both service disciplines in the first case. This is reflected in the term $d_1(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*))$. However, in the second case, the queue length process is different for each discipline. Under the exhaustive time-limited discipline, the server immediately leaves upon the queue becoming empty. Conversely, under the non-exhaustive time-limited discipline, the server remains at the queue and a sequence of idle and busy periods will follow until eventually the timer expires. The latter contribution to the queue-length process is represented in the term $d_2^{NE}(\mathbf{z})$.

Hence, $d_2^{NE}(\mathbf{z})$ reflects the p.g.f. of the number of customers at all queues at the end of a server visit process which runs for an exponential amount of time and which starts from an empty system. This function can be analyzed as follows. First, observe that the timer will interrupt the visit process either during an idle or a busy period. Second, observe that this process is regenerative in the sense that if the timer does not expire before the end of the first busy period, then the process starts like anew at that specific time instant. Let us denote by $I_i$ the length of an idle period at $Q_i$, by $BP_i$ the length of a busy period at $Q_i$ starting with a single customer, and by $T_i$ the exponential visit time of the server to $Q_i$. Then, we may write the following relation for $d_2^{NE}(\mathbf{z})$:

$$\begin{aligned} d_2^{NE}(\mathbf{z}) &= \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(T)}\mathbf{1}_{\{I_i>T_i\}}|\mathbf{N}_i(0)=\mathbf{n},\ N_{i,i}(0)=0] \\ &\quad + \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(T)}\mathbf{1}_{\{I_i<T_i, I_i+BP_i>T_i\}}|\mathbf{N}_i(0)=\mathbf{n},\ N_{i,i}(0)=0] \\ &\quad + \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(I_i+BP_i)}\mathbf{1}_{\{I_i+BP_i<T_i\}}|\mathbf{N}_i(0)=\mathbf{n},\ N_{i,i}(0)=0]\cdot d_2^{NE}(\mathbf{z}) \qquad (52) \\ &= \frac{\xi_i}{\lambda_i+\xi_i^*} + \frac{\lambda_i}{\lambda_i+\xi_i^*}\cdot\mathbb{E}[\mathbf{z}^{\mathbf{N}_i(T)}\mathbf{1}_{\{timer\}}|\mathbf{N}_i(0)=(n_1,\ldots,n_{i-1},1,n_{i+1},\ldots,n_M)] \\ &\quad + \frac{\lambda_i}{\lambda_i+\xi_i^*}\cdot\mu_i(\xi_i^*, r^i(\mathbf{z}))\cdot d_2^{NE}(\mathbf{z})\ , \qquad (53) \end{aligned}$$

where $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i(T)}\mathbf{1}_{\{timer\}}|\mathbf{N}_i(0) = \mathbf{n}]$ is provided in the analysis of the exhaustive time-limited discipline (see Prop. 2). Then, inserting this result of Prop. 2 and reorganizing the terms appropriately, we obtain:

$$\begin{aligned} d_2^{NE}(\mathbf{z}) &= \frac{\xi_i}{z - r^i(\mathbf{z})\cdot\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)}\times \\ &\left[\frac{(\lambda_i(1-z_i)+xi_i^*)(z-r^i(\mathbf{z})\cdot\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)) + \lambda_i\cdot z_i\cdot(1-\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*))(z_i-\mu_i(\xi_i^*, r^i(\mathbf{z}))}{(\lambda_i(1-z_i)+xi_i^*)(\lambda_i(1-\mu_i(\xi_i^*, r^i(\mathbf{z}))) + xi_i^*)}\right. \end{aligned}$$

where $\xi_i^* := \xi_i + \sum_{j\neq i}\lambda_j(1-z_j))$ . It can readily be verified that the latter expression is indeed equal to Eq. (51).

The above interpretation of the results suggests that similarly to the exhaustive time-limited discipline key relations for $\beta^i(\mathbf{z})$ may be found for any branching-property satisfying disciplines (e.g., the gated and the Bernouilli-type discipline) operating under an exponential time limit. Indeed for the gated time-limited discipline, we may readily find:

$$\text{Gated-TL:}\quad \beta^i(\mathbf{z}) = d_1^G(\mathbf{z})\cdot(\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^\bullet)) + d_2^G(\mathbf{z})\cdot\alpha^i(\mathbf{z}_i^\bullet), \qquad (55)$$

where $\alpha^i(\mathbf{z}_i^\bullet) := \alpha^i(z_1,\ldots,z_{i-1}, r^i(\mathbf{z})\cdot\tilde{S}_i(\xi_i^*), z_{i+1},\ldots,z_M)$, $d_1^G(\mathbf{z}) = d_1^E(\mathbf{z})$ and $d_2^G(\mathbf{z}) = d_2^E(\mathbf{z})$. This results follows by differentiating between the server departing due to having served all customers that were present at the start of the visit or due to the timer expiration. The former case readily gives the term $\alpha^i(\mathbf{z}_i^\bullet)$. In the latter case, the fact that each customer served was also present at the start of the visit leads to a straightforward analysis. By conditioning on the number of served customers and using that the LST of the service time of a successfully served customer equals $\tilde{S}(\xi_i + s)$, we obtain after some simple calculus the complementary part of Eq. (55).

Given the argumentation above, we strongly believe that these results carry over to any branching-property satisfying service discipline [7, 14] which is restricted by a timer. According to such as discipline, customers at $Q_i$ will effectively be replaced in an i.i.d. manner during the course of a server visit. Let us denote the corresponding p.g.f. which accounts for these replacements by $l_i(\mathbf{z})$. Then, we conclude this work with the following conjecture.

**Conjecture 1.** *For a single-server polling system with $Q_i$ operating under a branching-property satisfying service discipline with replacement p.g.f. $l_i(\boldsymbol{z})$ which is restricted by an exponentially distributed time limit, the queue-length evolution during a server visit to $Q_i$ can be described as follows:*

$$\beta^i(\boldsymbol{z}) = d(\boldsymbol{z}) \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^\star)) + \alpha^i(\boldsymbol{z}_i^\star), \tag{56}$$

*where* $\alpha^i(\boldsymbol{z}_i^\star) := \alpha^i(z_1, \dots, z_{i-1}, l_i(\boldsymbol{z}), z_{i+1}, \dots, z_M)$ *, and*

$$d(\boldsymbol{z}) = \frac{\xi_i}{z_i - r^i(\boldsymbol{z}) \cdot \tilde{S}_i(\lambda_i(1 - z_i) + \xi_i^*)} \cdot \frac{z_i \cdot (1 - \tilde{S}_i(\lambda_i(1 - z_i) + \xi_i^*))}{\lambda_i(1 - z_i) + \xi_i^*} \ . \tag{57}$$

# A Transient analysis of an M/G/1 during a busy period

In this section, we analyze the transient behavior of an M/G/1 queue during a busy period. We follow a similar approach as Cohen [13] used to study the transient behavior of the full queue-length process of the M/G/1 queue. To this end, we consider a single queue served by a single server. Customer arrive to the queue according to a Poisson process with rate $\lambda$. The service requirements $S$ of the customers are generally distributed with mean $b$.

Our interest is in the queue-length process during a busy period with some initial number of customers. Moreover, we keep track of the number of departures until time $t$. Therefore, similar to the transient transition probabilities $P_{hj}^{(n)}(t)$ that were defined in [13], we define the transient probabilities $R_{hj}^{(n)}(t)$ which specifically account for the fact that the system is nonempty from time 0 up to time $t$. More precisely, the transient probabilities $R_{hj}^{(n)}(t)$ are defined for $h, j, n = 1, 2, \ldots,$ and $t > 0$ as:

$$R_{hj}^{(n)}(t) := \mathbb{P}(z_n = j,\ r'_n < t,\ z_k > 0,\ 0 < k < n | z_0 = h)\ , \tag{58}$$

where it is assumed that at time $t = 0$ a new service starts. Notice that $R_{hj}^{(n)}(t)$ is only defined for $h, j \geq 1$. Our objective is to find an explicit expression for $\gamma_h(r, s, y)$ which is defined as:

$$\gamma_h(r, s, y) := \sum_{n=1}^{\infty} y^n \sum_{j=0}^{\infty} r^j \int_0^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h = 1, 2, \ldots\ . \tag{59}$$

From the definition of $R_{hj}^{(n)}(t)$, it follows immediately that:

$$\begin{aligned} R_{1j}^{(1)}(t) &= \int_{\tau=0}^{t} e^{-\lambda\tau} \frac{(\lambda\tau)^j}{j!} dS(\tau),\ j = 1, 2, \ldots, && (60)\\ R_{hj}^{(1)}(t) &= \int_{\tau=0}^{t} e^{-\lambda\tau} \frac{(\lambda\tau)^{j+1-h}}{(j+1-h)!} dS(\tau),\ j = h-1, h, \ldots,\ h = 2, 3, \ldots\ , && (61)\\ R_{hj}^{(1)}(t) &= 0,\ \text{otherwise}\ . && (62) \end{aligned}$$

Also, analogously to Eq. (4.20) of [13], we have the following recursive relation for $R_{hj}^{(n)}(t)$ for $t > 0$, $h, j = 1, 2, \ldots,\ n = 2, 3, \ldots,$

$$R_{hj}^{(n)}(t) = \sum_{l=1}^{\infty} \int_{u=0}^{t} R_{hl}^{(n-1)}(t-u) d_u R_{lj}^{(1)}(u)\ . \tag{63}$$

The following definitions will be used in the sequel:

$$\begin{aligned} \gamma_{hj}^{(n)}(s) &:= \int_0^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h, j, n = 1, 2, \ldots\ , && (64)\\ \gamma_h^{(n)}(r, s) &:= \sum_{j=1}^{\infty} r^j \gamma_{hj}^{(n)}(s),\ h, n = 1, 2, \ldots\ , && (65)\\ \gamma_{hj}(s, y) &:= \sum_{n=1}^{\infty} y^n \gamma_{hj}^{(n)}(s),\ h, j = 1, 2, \ldots\ , && (66)\\ \gamma_h(r, s, y) &:= \sum_{n=1}^{\infty} y^n \gamma_h^{(n)}(r, s),\ h = 1, 2, \ldots\ . && (67) \end{aligned}$$

As an immediate consequence of Eq. (63), we obtain the following result.

**Lemma 5.**

$$\gamma_h^{(n)}(r,s) = \sum_{l=1}^{\infty} \gamma_{hl}^{(n-1)}(s) \cdot \gamma_l^{(1)}(r,s), \;\; h = 1,2,\ldots, \;\; n = 2,3,\ldots, \;\; . \tag{68}$$

The final term in the right-hand side of Eq. (68), $\gamma_l^{(1)}(r,s)$, refers to the number of arrivals during a service time starting with $l$ customers. We have to distinguish between starting with one or with two or more customers, since in the former case the queue might be empty upon service completion and this situation should be excluded. A closed-form expression for this term is then given in the following lemma.

**Lemma 6.**

$$\gamma_1^{(1)}(r,s) = \tilde{S}\left(\lambda(1-r)+s\right) - \tilde{S}\left(\lambda+s\right) \;\; , \tag{69}$$

*and for* $h \geq 2$,

$$\gamma_h^{(1)}(r,s) = r^{h-1} \cdot \tilde{S}\left(\lambda(1-r)+s\right) \;\; . \tag{70}$$

*Proof.* Let us consider first the case $h \geq 2$:

$$\begin{aligned}
\gamma_h^{(1)}(r,s) &= \sum_{j=1}^{\infty} r^j \gamma_{hj}^{(1)}(s) && (71)\\
&= \sum_{j=h-1}^{\infty} r^j \gamma_{hj}^{(1)}(s) && (72)\\
&= \sum_{j=h-1}^{\infty} r^j \int_{t=0}^{\infty} e^{-st} dR_{hj}^{(1)}(t) && (73)\\
&= \int_{t=0}^{\infty} se^{-st} \sum_{j=h-1}^{\infty} r^j R_{hj}^{(1)}(t) dt && (74)\\
&= \int_{t=0}^{\infty} se^{-st} \int_{\tau=0}^{t} e^{-\lambda\tau} \sum_{j=h-1}^{\infty} r^{h-1} \cdot \frac{(r\lambda\tau)^{j+1-h}}{(j+1-h)!} \; dS(\tau) \; dt && (75)\\
&= r^{h-1} \int_{\tau=0}^{\infty} e^{-\lambda\tau(1-r)} \int_{t=\tau}^{\infty} s \cdot e^{-st} dt \; dS(\tau) && (76)\\
&= r^{h-1} \cdot \tilde{S}\left(\lambda(1-r)+s\right) \;\; . && (77)
\end{aligned}$$

In case $h = 1$, we should have at least one arrival before the first departure, otherwise the queue would become empty. Hence, in the derivation of $\gamma_1^{(1)}(r,s)$, we do not encounter the complete power series representation of the exponential function, so that the final expression will consist of two parts. More precisely,

$$
\begin{aligned}
\gamma_1^{(1)}(r,s) &= \sum_{j=1}^{\infty} r^j \gamma_{hj}^{(n)}(s) && (78)\\
&= \ldots = \int_{t=0}^{\infty} s e^{-st} \int_{\tau=0}^{t} e^{-\lambda\tau} \sum_{j=1}^{\infty} \frac{(r\lambda\tau)^j}{j!} dS(\tau)\ dt && (79)\\
&= \int_{\tau=0}^{\infty} e^{-\lambda\tau} \cdot (e^{-\lambda\tau r} - 1) \int_{t=\tau}^{\infty} s e^{-st} dt\ dS(\tau) && (80)\\
&= \tilde{S}\left(\lambda(1-r)+s\right) - \tilde{S}\left(\lambda+s\right)\ . && (81)
\end{aligned}
$$

□

Next, we are ready to present our main result of this section, i.e., a closed-form expression for $\gamma_h(r,s,y)$.

**Theorem 5.**

$$\gamma_h(r,s,y) = \frac{r}{r - y \cdot \tilde{S}\left(\lambda(1-r)+s\right)} \cdot \left(-\mu^h(s,y) + y \cdot \tilde{S}\left(\lambda(1-r)+s\right) \cdot r^{h-1}\right),\ h = 1, 2, \ldots\ , \tag{82}$$

*where $\mu(s,y)$ the smallest root of the function $x = y \cdot \tilde{S}\left(\lambda(1-x)+s\right)$ in $x$ with the absolute value smaller than one.*

*Proof.* Starting from the definition of $\gamma_h(r,s,y)$ and applying Lemmas 5 and 6, we obtain the following relations after some manipulations:

$$\gamma_1(r,s,y)\left(1 - \frac{y}{r} \cdot \tilde{S}\left(\lambda(1-r)+s\right)\right) = y \cdot \left(\tilde{S}\left(\lambda(1-r)+s\right) - \tilde{S}\left(\lambda+s\right) \cdot \left(1 + \gamma_{11}(s,y)\right)\right)\ , \tag{83}$$

$$\gamma_h(r,s,y)\left(1 - \frac{y}{r} \cdot \tilde{S}\left(\lambda(1-r)+s\right)\right) = y \cdot \left(\tilde{S}\left(\lambda(1-r)+s\right) \cdot r^{h-1} - \tilde{S}\left(\lambda+s\right) \cdot \gamma_{h1}(s,y)\right)\right)\ . \tag{84}$$

Denote by $\mu(s,y)$ the smallest root of the function $x = y \cdot \tilde{S}\left(\lambda(1-x)+s\right)$ in $x$ with the absolute value smaller than one. Since the functions $\gamma_h(r,s,y)$ should be analytic for $|r| \leq 1$, it follows that $\mu(s,y)$ is a zero of the right-hand side of the expressions above. Thus, we immediately obtain for $\gamma_{h1}(s,y)$:

$$
\begin{aligned}
\gamma_{11}(s,y) &= \frac{\mu(s,y) - y \cdot \tilde{S}\left(\lambda+s\right)}{y \cdot \tilde{S}\left(\lambda+s\right)}\ , && (85)\\
\gamma_{h1}(s,y) &= \frac{\mu^h(s,y)}{y \cdot \tilde{S}\left(\lambda+s\right)}\ ,\ h = 2, 3, \ldots\ . && (86)
\end{aligned}
$$

Notice that inserting $h = 1$ in the latter expression, which we denote by $(\gamma_{h1}(s,y))|_{h=1}$, shows that: $\gamma_{11}(s,y) + 1 = (\gamma_{h1}(s,y))|_{h=1}$. Finally, plugging these expressions into Eqs. (83) and (84) completes the proof. □

# B Proofs of results Section 3

In this section, we will give the proofs of the results of Sect. 3. For convenience, let us recall the following definitions for $t > 0$:

$$\begin{aligned}
p_{hk}^{(n)}(t) &:= \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n | x_0 = h), & h, k, n = 0, 1, \ldots, \\ 0, & \text{otherwise} . \end{cases} && (87)\\
P_{hj}^{(n)}(t) &:= \mathbb{P}(z_n = j,\ r'_n < t | z_0 = h),\ n = 1, 2, \ldots,\ h, j = 0, 1, \ldots\ , && (88)\\
F_k^{(0)}(t) &:= \mathbf{1}_{\{k=0\}}\mathbb{P}(A_i(t) = 0,\ I_i > t) + \mathbf{1}_{\{k\geq 1\}}\mathbb{P}(A_i(t) = k,\ I_i + S_i > t),\ k = 0, 1, \ldots\ , && (89)\\
F_k(t) &:= \mathbb{P}(A_i(t) = k,\ S_i > t),\ k = 0, 1, \ldots,\ . && (90)
\end{aligned}$$

## B.1 Proof of Lemma 1

The proof of the lemma is carried out as follows. First, we rewrite the event $D(t) = n$ and use the assumption that at time 0 the 0-th customer departed from the queue, so that we obtain Eq. (92). Next, we condition on the number of customers present at the $n$th departure, $z_n$, and on the time this departure occurs, $r'_n$, which leads to Eq. (93). Finally, observing that $r_{n+1},\ n = 0, 1, \ldots,$ depends in fact only $r_n$ and $z_n$, using that the arrival process is stationary and applying the definitions of $F_k^{(0)}(t),\ F_k(t)$ and $P_{hj}^{(n)}(t)$ provides us with Eq. (94).

$$\begin{aligned}
p_{hk}^{(n)}(t) &:= \mathbb{P}(x_t = k, D(t) = n | x_0 = h) && (91)\\
&= \mathbb{P}(x_t = k, r'_n \leq t, r'_{n+1} > t | z_0 = h) && (92)\\
&= \int_{u=0}^{t} \sum_{j=0}^{k} \mathbb{P}(x_t = k,\ r'_{n+1} > t |\ r'_n = u,\ z_0 = h,\ z_n = j)\\
&\quad \times d_u\mathbb{P}(\ r'_n \leq u,\ z_n = j | z_0 = h) && (93)\\
&= \int_{u=0}^{t} F_k^{(0)}(t-u) dP_{h0}^{(n)}(u) + \sum_{j=1}^{k} \int_{u=0}^{t} F_{k-j}(t-u) dP_{hj}^{(n)}(u)\ . && (94)
\end{aligned}$$

Let us define the following LSTs.

$$\begin{aligned}
\tilde{F}_k^{(0)}(s) &:= \int_{0-}^{\infty} e^{-st} dF_k^{(0)}(t),\ k = 0, 1, \ldots\ , && (95)\\
\tilde{F}_k(s) &:= \int_{0-}^{\infty} e^{-st} dF_k(t),\ k = 0, 1, \ldots\ , && (96)\\
\pi_{hj}^{(n)}(s) &:= \int_{0-}^{\infty} e^{-st} dP_{hj}^{(n)}(t),\ n = 1, 2, \ldots,\ h, j = 0, 1, \ldots\ . && (97)
\end{aligned}$$

Then, we may present the following result as an immediate consequence of Lemma 1:

**Corollary 1.**

$$\int_{t=0-}^{\infty} e^{-st} dp_{hk}^{(n)}(t) = \tilde{F}_k^{(0)}(s)\pi_{h0}^{(n)}(s) + \sum_{j=1}^{k} \tilde{F}_k^{(j)}(s)\pi_{hj}^{(n)}(s)\ . \qquad (98)$$

### B.2 Proof of Lemma 2

Before we get to the actual proof of Lemma 2, we present another lemma. Let us introduce the auxiliary functions $G_i^{(0)}(r,s)$ and $G_i(r,s)$. These functions refer to the number of customers that arrive to the system during a period which starts at a service completion instant and ends at a timer expiration which occurs before a next service is completed. More specifically, the function $G_i^{(0)}(r,s)$ refers to the case with zero customers present after a service completion, while $G_i(r,s)$ refers to the case with a strictly positive number of customers present at a service completion instant.

**Lemma 7.**

$$\begin{aligned} G_i^{(0)}(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k^{(0)}(s) \\ &= \frac{s}{\lambda_i(1-r)+s} \cdot \frac{\lambda_i(1-r\cdot\tilde{S}_i(\lambda_i(1-r)+s))+s}{\lambda_i+s} , \qquad (99) \\ G_i(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k(s) \\ &= \frac{s}{\lambda_i(1-r)+s} \cdot \Big(1-\tilde{S}_i(\lambda_i(1-r)+s)\Big) . \qquad (100) \end{aligned}$$

*Proof.* First, we will prove the expression for $G_i^{(0)}(r,s)$. We separate the terms for $k=0$ and $k \geq 1$, insert the expression for $\tilde{F}_k^{(0)}(s)$ and perform some simple calculations yielding Eq. (102). Next, we condition on the interarrival time (for the case $k \geq 1$) and use the fact that for a given time $t$ the events $\{A_i(t)=k\}$ and $\{S_i > t\}$ are independent. The final expression, Eq. (103), then readily follows from the Poisson arrival assumption and some simple manipulations.

$$\begin{aligned} G_i^{(0)}(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k^{(0)}(s) \qquad (101) \\ &= s\cdot\int_{t=0}^{\infty} e^{-st}\mathbb{P}(A_i(t)=0)dt \\ &\quad + r\cdot\sum_{k=1}^{\infty} r^{k-1}\cdot s\cdot\int_{t=0}^{\infty} e^{-st}\mathbb{P}(A_i(t)=k,\ I_i+S_i>t)dt \qquad (102) \\ &= \frac{s}{\lambda_i(1-r)+s}\cdot\frac{\lambda_i(1-r\cdot\tilde{S}_i(\lambda_i(1-r)+s))+s}{\lambda_i+s} . \qquad (103) \end{aligned}$$

Analogously, we find for $G_i(r,s)$:

$$\begin{aligned} G_i(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k(s) \qquad (104) \\ &= \sum_{k=0}^{\infty} r^k\cdot s\int_{t=0}^{\infty} e^{-st}\mathbb{P}(A_i(t)=k,\ S_i>t)dt \qquad (105) \\ &= \frac{s}{\lambda_i(1-r)+s}\cdot\Big(1-\tilde{S}_i(\lambda_i(1-r)+s)\Big) . \qquad (106) \end{aligned}$$

□

Let us give several definitions which will be used in the proof of Lemma 2:

$$\pi_h^{(n)}(r,s) := \sum_{j=0}^{\infty} r^j \pi_{hj}^{(n)}(s),\ h=0,1,\dots,\ n=1,2,\dots\ , \tag{107}$$

$$\pi_{h0}(s,y) := \sum_{n=1}^{\infty} y^n \pi_{h0}^{(n)}(s),\ h=0,1,\dots\ , \tag{108}$$

$$\pi_h(r,s,y) := \sum_{n=1}^{\infty} y^n \pi_h^{(n)}(r,s),\ h=0,1,\dots\ . \tag{109}$$

*Proof of Lemma 2.* The proof of Lemma 2 consists in fact of three main steps. In the first step, we substitute the result of Corollary 1 into Eq. (110) leading to Eq. (111). Next, we work out the generating function with respect to the number of customers at the end of a visit. After some manipulations and using the definitions of $G_i^{(0)}(r,s)$, $G_i(r,s)$, $\pi_{h0}^{(n)}(s)$ and $\pi_h^{(n)}(r,s)$, we arrive at Eq. (112). In the final step, we use the definitions of $\pi_h(r,s,y)$ and $\pi_{h0}(s,y)$ and insert the explicit expressions for $G_i^{(0)}(r,s)$ and $G_i(r,s)$ which were derived in Lemma 7.

$$\sum_{n=1}^{\infty} y^n \sum_{k=0}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dp_{hk}^{(n)}(t) \tag{110}$$

$$= \sum_{n=1}^{\infty} y^n \sum_{k=0}^{\infty} r^k \left( \tilde{F}_k^{(0)}(s)\pi_{h0}^{(n)}(s) + \sum_{j=1}^{k} \tilde{F}_k^{(j)}(s)\pi_{hj}^{(n)}(s) \right) \tag{111}$$

$$= \sum_{n=1}^{\infty} y^n \left( G_i^{(0)}(r,s)\cdot \pi_{h0}^{(n)}(s) + G_i(r,s)\left(\pi_h^{(n)}(r,s) - \pi_{h0}^{(n)}(s)\right)\right) \tag{112}$$

$$= \frac{s}{\lambda_i(1-r)+s} \cdot \frac{\lambda_i(1-r\cdot\tilde{S}_i(\lambda_i(1-r)+s))+s}{\lambda_i+s} \cdot \pi_{h0}(s,y)$$
$$+ \frac{s}{\lambda_i(1-r)+s} \cdot (1-\tilde{S}_i(\lambda_i(1-r)+s)) \cdot (\pi_h(r,s,y) - \pi_{h0}(s,y))\ ,\ h=0,1,\dots . \tag{113}$$

□

### B.3 Proof of Theorem 2

We prove the expression for $\beta^i(\mathbf{z})$ as given in Theorem 2 by first deriving the conditional p.g.f. $\beta^i_{\mathbf{n}}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}|\mathbf{N}_i^s = \mathbf{n}]$ and then unconditioning on $\mathbf{N}_i^s$, the number of customers present at the start of a visit to $Q_i$. For convenience, let us define $\xi_i^*$ as follows.

$$\xi_i^* := \xi_i + \sum_{j\neq i} \lambda_j(1-z_j)\ . \tag{114}$$

Next, $\beta^i_{\mathbf{n}}(\mathbf{z})$ can be expressed as follows.

**Lemma 8.**

$$\beta^i_{\boldsymbol{n}}(\boldsymbol{z}) \;=\; \frac{\xi_i}{\xi_i^*} \cdot \Big( G_i^{(0)}(z_i, \xi_i^*) \cdot \big(\pi_{n_i,0}(\xi_i^*, r^i(\boldsymbol{z})) + 1_{\{n_i=0\}}\big) \tag{115}$$

$$+\; G_i(z_i, \xi_i^*) \cdot \big(\pi_{n_i}(z_i, \xi_i^*, r^i(\boldsymbol{z})) + z_i^{n_i} - \pi_{n_i,0}(\xi_i^*, r^i(\boldsymbol{z})) - 1_{\{n_i=0\}}\big)\Big) \cdot \prod_{j\neq i} z_j^{n_j} \;.$$

*Proof.* Let $A_{i,j}(t)$ denote the number of arrivals to $Q_j$ (both external and internal arrivals) during a visit to $Q_i$. Recall further that $D(t)$ denotes the number of departures at $Q_i$ from time 0 to $t$. Starting from the definition of the p.g.f., we condition on the timer $T_i$ and introduce the number of departures from $Q_i$ until time $t$, $D(t)$.

$$\beta^i_{\mathbf{n}}(\mathbf{z}) = \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1} \cdots z_M^{m_M} \mathbb{P}(\mathbf{N}_i^e = \mathbf{m} | \mathbf{N}_i^s = \mathbf{n}) \tag{116}$$

$$= \int_0^\infty \xi_i e^{-\xi_i t} \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1} \cdots z_M^{m_M} \sum_n \mathbb{P}(\mathbf{N}_i(t) = \mathbf{m},\ D(t) = n | \mathbf{N}_i(0) = \mathbf{n}) dt \tag{117}$$

After some simple rearrangements and using that given $t$ and $D(t)$ the queue-length process at $Q_i$ is independent of the aggregate arrival process to the other queues, we obtain the following:

$$\int_0^\infty \xi_i e^{-\xi_i t} \sum_n \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1-n_1} \cdots z_M^{m_M-n_M}$$
$$\times\, \mathbb{P}(\{A_{i,j}(t) = m_j - n_j,\ \forall_{j\neq i}\} | D(t) = n,\ \mathbf{N}_i(0) = \mathbf{n})$$
$$\times \sum_{m_i} z_i^{m_i} \mathbb{P}(N_{i,i}(t) = m_i | D(t) = n,\ \mathbf{N}_i(0) = \mathbf{n})\ \mathbb{P}(D(t) = n | \mathbf{N}_i(0) = \mathbf{n}) dt \cdot \prod_{j\neq i} z_j^{n_j} \tag{118}$$

These aggregate arrivals to $Q_j$, $j \neq i$, can be decomposed in two independent parts, viz., a first part referring to external arrivals at each queue and a second part referring to customers that were served at $Q_i$ and routed to some other queue. The latter is represented by the term $(r^i(\mathbf{z}))^n$. Also noting that $N_{i,i}(t)$ depends only on $\mathbf{N}_i(0)$ through $N_{i,i}(0)$, we retrieve $p^{(n)}_{n_i m_i}(t)$ and eventually find that $\beta^i_{\mathbf{n}}(\mathbf{z})$ equals:

$$\int_0^\infty \xi_i e^{-\xi_i^* t} \sum_{n=0}^{\infty} \sum_{m_i=0}^{\infty} z_i^{m_i} (r^i(\mathbf{z}))^n p^{(n)}_{n_i m_i}(t) dt \cdot \prod_{j\neq i} z_j^{n_j} \tag{119}$$

Then, we can apply Lemma 2 for $n \geq 1$, while for $n = 0$ we use:

$$\sum_{m_i=0}^{\infty} z_i^{m_i} \int_0^\infty \xi_i e^{-\xi_i^* t} p^{(0)}_{n_i m_i}(t) dt \tag{120}$$

$$= \; \mathbf{1}_{\{n_i=0\}} \cdot \sum_{m_i=0}^{\infty} z_i^{m_i} \int_0^\infty \xi_i e^{-\xi_i^* t} \mathbb{P}(A_i(t) = m_i,\ I_i + S_i > t) dt$$

$$+\; \mathbf{1}_{\{n_i\geq 1\}} \cdot \sum_{m_i=0}^{\infty} z_i^{m_i} \int_0^\infty \xi_i e^{-\xi_i^* t} \mathbb{P}(A_i(t) = m_i - n_i,\ S_i > t) dt \tag{121}$$

$$= \; \frac{\xi_i}{\xi_i^*} \cdot \Big( \mathbf{1}_{\{n_i=0\}} \cdot G_i^{(0)}(z_i, \xi_i^*) + \mathbf{1}_{\{n_i\geq 1\}} \cdot z_i^{n_i} \cdot G_i(z_i, \xi_i^*) \Big) \;. \tag{122}$$

This leads after some manipulations to the final expression for $\beta^i_{\mathbf{n}}(\mathbf{z})$:

$$\beta^i_{\mathbf{n}}(\mathbf{z}) = \frac{\xi_i}{\xi_i^*} \cdot \Big( G_i^{(0)}(z_i, \xi_i^*) \cdot \big(\pi_{n_i,0}(\xi_i^*, r^i(\mathbf{z})) + 1_{\{n_i=0\}}\big) \\ + \; G_i(z_i, \xi_i^*) \cdot \big(\pi_{n_i}(z_i, \xi_i^*, r^i(\mathbf{z})) + z_i^{n_i} - \pi_{n_i,0}(\xi_i^*, r^i(\mathbf{z})) - 1_{\{n_i=0\}}\big)\Big) \cdot \prod_{j \neq i} z_j^{n_j} \;. \tag{123}$$

□

*Proof of Theorem 2.* Essentially, the proof follows immediately by unconditioning $\beta^i_{\mathbf{n}}(\mathbf{z})$ on the state $\mathbf{n} = (n_1, \ldots, n_M)$ at the start of the visit. The result of this operation is shown below. Equation (125) follows by substitution of Eq. (115) into the definition of $\beta^i(\mathbf{z})$. We note that the final expression, Eq. (126), follows from inserting the explicit expressions for $G_i^{(0)}(r,s)$ and $G_i(r,s)$ (see Lemma 7), inserting the expressions for $\pi_h(z_i, \xi_i^*, r^i(\mathbf{z}))$ and $\pi_{h0}(\xi_i^*, r^i(\mathbf{z}))$, $h \geq 0$, which are given in Eqs. (15), (20) and (21), and some simple manipulations.

$$\begin{aligned}
\beta^i(\mathbf{z}) &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \beta^i_{\mathbf{n}}(\mathbf{z}) \mathbb{P}(\mathbf{N}_i^s = \mathbf{n}) && (124)\\
&= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \mathbb{P}(\mathbf{N}_i^s = \mathbf{n}) \cdot \prod_{j\neq i} z_j^{n_j} \cdot \frac{\xi_i}{\xi_i^*} \cdot \Big( G_i^{(0)}(z_i, \xi_i^*) \cdot \big(\pi_{n_i,0}(\xi_i^*, r^i(\mathbf{z})) + 1_{\{n_i=0\}}\big) \\
&\quad + G_i(z_i, \xi_i^*) \cdot \big(\pi_{n_i}(z_i, \xi_i^*, r^i(\mathbf{z})) + z_i^{n_i} - \pi_{n_i,0}(\xi_i^*, r^i(\mathbf{z})) - 1_{\{n_i=0\}}\big)\Big) && (125)\\
&= \frac{\xi_i}{z_i - r^i(\mathbf{z}) \cdot \tilde{S}_i(\xi_i + \sum_j \lambda_j (1 - z_j))} \\
&\quad \times \left\{ \frac{\tilde{S}_i(\xi_i + \sum_j \lambda_j(1-z_j)) \cdot (z_i - r^i(\mathbf{z}))}{\lambda_i(1 - \mu(\xi_i, r^i(\mathbf{z}))) + \xi_i^*} \cdot \alpha^i(\mathbf{z}_i^*) \right. \\
&\quad \left. + \frac{(1 - \tilde{S}_i(\xi_i + \sum_j \lambda_j(1 - z_j))) \cdot z_i}{\lambda_i(1 - z_i) + \xi_i^*} \cdot \alpha^i(\mathbf{z}) \right\} , && (126)
\end{aligned}$$

where $\alpha^i(\mathbf{z}_i^*) := \mathbb{E}[z_1^{N_1^s} \cdots \mu(\xi_i^*, z)^{N_i^s} \cdots z_M^{N_M^s}]$ and $\mu(\xi_i^*, r^i(\mathbf{z}))$ is the root $x$ with the smallest absolute value less than one of $x = r^i(\mathbf{z}) \cdot \tilde{S}_i(\xi_i^* + \lambda_i(1 - x))$. □

# C Proofs of results Section 4

In this section, we will give the proofs of the results of Sect. 4. For convenience, let us recall the following definitions for $t > 0$:

$$\begin{aligned}
q_{hk}^{(n)}(t) &:= \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n,\ x_v > 0,\ 0 < v < t | x_0 = h), & n = 0, 1, \ldots, \quad h, k = 1, 2, \ldots, \\ 0, & \text{otherwise}, \end{cases} && (127)\\
R_{hj}^{(n)}(t) &:= \mathbb{P}(z_n = j,\ r_n' < t,\ z_k > 0,\ 0 < k < n | z_0 = h),\ h, j, n = 1, 2, \ldots\ , && (128)\\
F_k(t) &:= \mathbb{P}(A_i(t) = k,\ S_i > t),\ k = 0, 1, \ldots\ . && (129)
\end{aligned}$$

## C.1 Proof of Proposition 1

The first observation is that each customer at $Q_j,\ j \neq i$, will still be present at the end of the visit, which is accounted for in the term $\prod_{j\neq i} z_j^{n_j}$. Second, each customer present at the start of the visit at $Q_i$ will effectively be replaced by a random population during the course of the visit in an identical fashion. In particular, the size of this population is given by $\mu_i(\xi_i^*, r^i(\mathbf{z}))$. To see this, recall that $\mu_i(s,1)$ equals the LST with parameter $s$ of the length of the busy period at $Q_i$. The term $\xi_i^*$ in $\mu_i(\xi_i^*, r^i(\mathbf{z}))$ accounts for the exogenous arrivals to the other queues in the system during a busy period which ends before the timer expires. Similarly, the term $r^i(\mathbf{z})$ in $\mu_i(\xi_i^*, r^i(\mathbf{z}))$ accounts for the internal arrivals to the other queues (from $Q_i$) during this period. As initially there are $n_i$ identical customers present at $Q_i$, this leads to $n_i$ independent contributions which are recognized in the power of $\mu_i(\xi_i^*, r^i(\mathbf{z}))$.

## C.2 Proof of Lemma 3

Lemma 3 is readily proven by using similar arguments as in the proof of Lemma 1:

$$\begin{aligned} q_{hk}^{(n)}(t) &= \mathbb{P}(x_t = k,\ D(t) = n,\ x_v > 0,\ 0 < v < t | x_0 = h) && (130)\\ &= \mathbb{P}(x_t = k,\ r_n' \leq t,\ r_{n+1}' > t,\ x_v > 0,\ 0 < v < t | z_0 = h) && (131)\\ &= \int_{u=0}^{t} \sum_{j=1}^{k} \mathbb{P}(x_t = k,\ r_{n+1}' > t |\ r_n' = u,\ z_0 = h,\ z_m > 0,\ 0 \leq m \leq n,\ z_n = j) && (132)\\ &\qquad\qquad \times d_u \mathbb{P}(\ r_n' \leq u,\ z_n = j,\ z_m > 0,\ 0 < m < n | z_0 = h)\\ &= \int_{u=0}^{t} \sum_{j=1}^{k} F_k^{(j)}(t-u) dR_{hj}^{(n)}(u)\ . && (133) \end{aligned}$$

Let us define the following LSTs.

$$\begin{aligned} \tilde{F}_k(s) &:= \int_{0-}^{\infty} e^{-st} dF_k(t),\ k = 0, 1, \dots\ , && (134)\\ \gamma_{hj}^{(n)}(s) &:= \int_{0-}^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h, j, n = 1, 2, \dots\ . && (135) \end{aligned}$$

Then, a direct consequence of Lemma 3 is:

**Corollary 2.**

$$\int_{t=0}^{\infty} e^{-st} dq_{hk}^{(n)}(t) = \sum_{j=1}^{k} \gamma_{hj}^{(n)}(s) \tilde{F}_k(s),\ h, k, n = 1, 2, \dots\ . \qquad (136)$$

## C.3 Proof of Lemma 4

Let us give several definitions which will be used in the proof of Lemma 4:

$$\begin{aligned} \gamma_h^{(n)}(r,s) &:= \sum_{j=0}^{\infty} r^j \gamma_{hj}^{(n)}(s),\ h, n = 1, 2, \dots\ , && (137)\\ \gamma_h(r,s,y) &:= \sum_{n=1}^{\infty} y^n \gamma_h^{(n)}(r,s),\ h = 1, 2, \dots\ . && (138) \end{aligned}$$

*Proof of Lemma 4.* The proof exists of three consecutive steps similar to the proof of Lemma 2. First, we substitute Eq. (136) into Eq. (139). Next, using the definitions of $\gamma_h^{(n)}(r,s)$ and $G_i(r,s)$ immediately yields Eq. (141). The final step follows from the definition of $\gamma_h(r,s,y)$ and the substitution of the explicit expression for $G_i(r,s)$ (see Lemma 7).

$$\begin{aligned}
&\sum_{n=1}^{\infty} y^n \sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(n)}(t) && (139)\\
&= \sum_{n=1}^{\infty} y^n \sum_{k=1}^{\infty} r^k \sum_{j=1}^{k} \gamma_{hj}^{(n)}(s)\tilde{F}_k^{(j)}(s) && (140)\\
&= \sum_{n=1}^{\infty} y^n \gamma_h^{(n)}(r,s) G_i(r,s) && (141)\\
&= \gamma_h(r,s,y) \cdot \frac{s}{\lambda_i(1-r)+s} \cdot (1-\tilde{S}_i(\lambda_i(1-r)+s)) \ . && (142)
\end{aligned}$$

□

## C.4 Proof of Proposition 2

As a preliminary to proving Proposition 2, we present the following result for the special case of $D(t) = 0$, i.e., no departures occur before the timer expires.

**Lemma 9.**

$$\sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(0)}(t) = r^h \cdot \frac{s}{\lambda_i(1-r)+s} \cdot (1-\tilde{S}_i(\lambda_i(1-r)+s)), \ h = 1, 2, \dots \ . \qquad (143)$$

*Proof.* Elaborating on the definition of $q_{hk}^{(0)}(t)$, we may obtain after some simple manipulations Eq. (145). Equation (146) then follows directly from the earlier derivation of $G_i(r,s)$ (see Eq. (105)).

$$\begin{aligned}
&\sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(0)}(t) && (144)\\
&= r^h \cdot \int_{t=0}^{\infty} s e^{-st} \sum_{k=h}^{\infty} r^{k-h} \mathbb{P}(A_i(t) = k-h, \ S_i > t) dt && (145)\\
&= r^h \cdot \frac{s}{\lambda_i(1-r)+s} \cdot (1-\tilde{S}_i(\lambda_i(1-r)+s)) \ . && (146)
\end{aligned}$$

□

*Proof of Proposition 2.* Let $A_{i,j}(t)$ denote the number of arrivals to $Q_j$ (both external and internal arrivals) during a visit to $Q_i$. Recall further that $D(t)$ denotes the number of departures at $Q_i$ from time 0 to $t$. Starting from the definition of the p.g.f., we condition on the timer $T_i$ and introduce the number of departures from $Q_i$ until time $t$, $D(t)$.

$$\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s=\mathbf{n}] \tag{147}$$

$$= \sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty} z_1^{m_1}\cdots z_M^{m_M}\mathbb{P}(\mathbf{N}_i^e=\mathbf{m},\ \{timer\}|\mathbf{N}_i^s=\mathbf{n}) \tag{148}$$

$$= \int_0^{\infty}\xi_i e^{-\xi_i t}\sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty} z_1^{m_1}\cdots z_M^{m_M}$$

$$\times\sum_n \mathbb{P}(\mathbf{N}_i(t)=\mathbf{m},\ \{timer\},\ D(t)=n|\mathbf{N}_i(0)=\mathbf{n})dt \tag{149}$$

Using that given $t$ and $D(t)$ the queue-length process at $Q_i$ is independent of the aggregate arrival process to the other queues and working out the event $\{timer\}$, we obtain:

$$\int_0^{\infty}\xi_i e^{-\xi_i t}\sum_n\sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty} z_1^{m_1-n_1}\cdots z_M^{m_M-n_M}$$

$$\times\mathbb{P}(\{A_{i,j}(t)=m_j-n_j,\ \forall_{j\neq i}\}|D(t)=n,\ \mathbf{N}_i(0)=\mathbf{n})$$

$$\times\sum_{m_i} z_i^{m_i}\mathbb{P}(N_{i,i}(t)=m_i,\quad N_{i,i}(v)>0,\ 0<v<t|D(t)=n,\ \mathbf{N}_i(0)=\mathbf{n}) \tag{150}$$

$$\times\mathbb{P}(D(t)=n|\mathbf{N}_i(0)=\mathbf{n})dt\cdot\prod_{j\neq i} z_j^{n_j} \tag{151}$$

Exactly following the same reasoning that lead to Eq. (119), we have that $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s=\mathbf{n}]$ equals:

$$\int_{t=0}^{\infty}\xi_i e^{-\xi_i^* t}\sum_{n=0}^{\infty}\sum_{m_i=1}^{\infty} z_i^{m_i} r^i(\mathbf{z})^n q_{n_i m_i}^{(n)}(t)dt\cdot\prod_{j\neq i} z_j^{n_j} \tag{152}$$

Then, we may apply Lemma 4 for $n\geq 1$ and Lemma 9 for $n=0$. This leads after substituting the explicit expressions for $G_i(r,s)$ (see Lemma 7) and $\gamma_h(r,s,y)$ (see Eq. (15)) and performing some simple manipulations to the final expression, viz.:

$$\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s=\mathbf{n}]$$

$$= \frac{\xi_i\cdot z_i\cdot(1-\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*))}{[\lambda_i(1-z_i)+\xi_i^*]\cdot[z_i-r^i(\mathbf{z})\cdot\tilde{S}_i(\lambda_i(1-z_i)+\xi_i^*)]}\cdot\left(z_i^{n_i}-\mu^{n_i}(\xi_i^*,r^i(\mathbf{z}))\right)\cdot\prod_{j\neq i} z_j^{n_j}\ . \tag{153}$$

□

## C.5 Proof of Theorem 4

The final result for $\beta^i(\mathbf{z})$ is obtained by unconditioning the conditional p.g.f.'s of the previous propositions and then merging these outcomes. Let us define $\beta_e^i(\mathbf{z}):=\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}]$, $\beta_t^i(\mathbf{z}):=\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}]$ and $\alpha^i(\mathbf{z}_i^*):=\alpha^i(z_1,\ldots,z_{i-1},\mu_i(\xi_i^*,r^i(\mathbf{z})),z_{i+1},\ldots,z_M)$. First, $\beta_e^i(\mathbf{z})$ and $\beta_e^i(\mathbf{z})$ are given in the following two lemmas which are immediate from unconditioning the expressions in Propositions 1 and 2.

**Lemma 10.**

$$\beta_e^i(\boldsymbol{z}) = \alpha^i(\boldsymbol{z}_i^*) \ . \tag{154}$$

**Lemma 11.**

$$\beta_t^i(\boldsymbol{z}) = \frac{\xi_i \cdot z_i \cdot (1 - \tilde{S}_i(\lambda_i(1-z_i) + \xi_i^*))}{(\lambda_i(1-z_i) + \xi_i^*)(z_i - r^i(\boldsymbol{z}) \cdot \tilde{S}_i(\lambda_i(1-z_i) + \xi_i^*))} \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) \ . \tag{155}$$

*Proof of Theorem 4.* The proof follows directly from the two final lemmas above:

$$\begin{aligned} \beta^i(\mathbf{z}) &= \beta_e^i(\mathbf{z}) + \beta_t^i(\mathbf{z}) && (156)\\ &= \left(1 - \frac{\xi_i \cdot z_i \cdot (1 - \tilde{S}_i(\lambda_i(1-z_i) + \xi_i^*))}{(\lambda_i(1-z_i) + \xi_i^*)(z_i - r^i(\mathbf{z}) \cdot \tilde{S}_i(\lambda_i(1-z_i) + \xi_i^*))}\right) \cdot \alpha^i(\mathbf{z}_i^*) \\ &\quad + \frac{\xi_i \cdot z_i \cdot (1 - \tilde{S}_i(\lambda_i(1-z_i) + \xi_i^*))}{(\lambda_i(1-z_i) + \xi_i^*)(z_i - r^i(\mathbf{z}) \cdot \tilde{S}_i(\lambda_i(1-z_i) + \xi_i^*))} \cdot \alpha^i(\mathbf{z}) \ . && (157) \end{aligned}$$

□

# D Proof of Theorem ??

## D.1 Proof: Preliminaries and stochastic monotonicity

We will prove the theorem adopting the approach of Fricker and Jaïbi [15]. To this end, we will often stick to their notation whenever it does not lead to ambiguity. We should emphasize that the authors of [15] considered only work-conserving service disciplines. The exhaustive time-limited (ETL) discipline allows for preemption of service and thus is definitely not work conserving.

To circumvent this problem, one could modify the service requirements by introducing *effective* service times which account also for the unsuccessful service attempts. Since the time limit is assumed exponentially distributed, this effective service time could readily be seen to be a geometric sum of specific individual service attempt durations. However, if one would do so, then it may lead to conflicts with the behavior in the original system. For instance, in case of deterministic service times with mean $D$, it could happen that in the modified system the server works on the same customer without interruption for a period of time longer than $D$.

The general service disciplines considered in [15] should satisfy four properties. Property 1 and 3 refer to the independence of the service discipline on the history of the service process and on the independence of the customer selection. These properties are readily seen to be satisfied for the ETL discipline. Property 2 deals with the work conservation and is not satisfied for this discipline. Finally, Property 4 is the so-called stochastic monotonicity property and is defined as follows [15]: "*As the queue size grows, the number of customers served during one stage (visit) grows stochastically, but such that the number of customers left at the end of the stage (visit) grows stochastically as well.*" This latter property plays in crucial role in the proof.

Let us w.l.o.g. consider an arbitrary queue in the polling system. First, we define the modified service times of a customer, $(\sigma^m)_m$, with mean $\sigma$ being distributed as $\min(S_m, T)$, where $S_m$ is the original service time and $T$ is the duration of the exponential timer. That is, the modified service times can be seen as the duration of a service attempt (which can either be successful or interrupted). For non-preemptive service disciplines, the number of customers taken into service is equal to the number of customers served during a visit. However, this is not always true for the preemptive discipline that we consider here. Hence, we will define also the following quantities for a visit with $x$ customers present at the start (i.e., $t = 0$):

- $f^+(x)$: the number of customers that is taken into service during the visit;
- $f^-(x)$: the number of customers that is actually served during the visit;
- $v(x)$: the duration of the visit;
- $\phi(x)$: the number of customers at the end of the visit .

We note that the ETL discipline is not work conserving, since work is created due to preemptions. However, during the course of a visit the server is always working and does not idle. Thus, we may write the following relations between $f^+(x)$, $f^-(x)$, $v(x)$ and $\phi(x)$:

$$\begin{aligned} v(x) &= \sum_{m=1}^{f^+(x)} \sigma^m \ , && (158) \\ \phi(x) &= x - f^-(x) + N(0, v(x)] \ , && (159) \end{aligned}$$

with $f^+(0) = f^-(0) = v(0) = 0$.

Let us next briefly recall the definitions of $\leq$-monotonicity and $\leq_d$-monotonicity as given in [15]:

**Definition 1.** *($\leq$-monotonicity)*
*A real function $h$ defined on $\mathbb{R}^n$ is called $\leq$-monotone when:*

$$x \leq y \Rightarrow h(x) \leq h(y) \ . \tag{160}$$

**Definition 2.** *($\leq_d$-monotonicity)*
*Two (cumulative) distributions functions $P_1$ and $P_2$ on $\mathbb{R}^n$ satisfy $P_1 \leq_d P_2$ when:*

$$\int h \ dP_1 \leq \int h \ dP_2 \ , \tag{161}$$

*for any $\leq$-monotone function $h$ such that the integrals are well defined.*
*Two random vectors $X_1$ and $X_2$ satisfy $X_1 \leq_d X_2$ if their distribution satisfy $P_1 \leq_d P_2$.*

Hence, the monotonicity property for the ETL discipline is that $(f^+(x), f^-(x), \phi(x))$ is $\leq_d$-monotone in $x$. It follows immediately from Eq. (158) that $\leq_d$-monotonicity of $f^+(x)$ implies $\leq_d$-monotonicity of $v(x)$, but not that $\leq_d$-monotonicity of $f^-(x)$ implies $\phi(x)$. Note that the latter may seem true, but observe that so far no assumptions are made on the service disciplines (e.g., for general threshold disciplines this statement is typically false; see also Remark 2 in [15]).

Next, we embed the queue into the polling system. Let the $n$th visit to the queue start at stopping time $T_n$ with $Q_n$ customers waiting. Define the following quantities:

- $F_n^+$: the number of customers that is taken into service during visit $n$;
- $F_n^-$: the number of customers that is actually served during visit $n$;

- $V_n$: the duration of visit $n$;
- $\Phi_n$: the number of customers at the end of visit $n$;

Let us introduce the tuple $(f^+, f^-, v, \phi)$. It can readily be argued (cf. [15, p.215]) that for each $n$:

$$(F_n^+, F_n^-, V_n, \Phi_n) =_d (f^+(Q_n), f^-(Q_n), v(Q_n), \phi(Q_n) \ . \tag{162}$$

The service discipline is then characterized by the distribution of $(f^+, f^-, v, \phi)$ and we refer to it as discipline $f$. Along the single-queue equations, Eqs. (158) and (159), we find that for any $n$, $V_n$ and $\Phi_n$ are related to $F_n^+$ and $F_n^-$ as follows.

$$\begin{aligned} V_n &= \sum_{i=D_n+1}^{D_n+F_n^+} \sigma^i \ , & (163)\\ \Phi_n &= Q_n - F_n^- + N(T_n, T_n + V_n] \ , & (164) \end{aligned}$$

where $D_n$ denotes the number of customers served up to $T_n$ and $N(a, b]$ the number of arrivals to the queue during the interval $(a, b]$. Using Wald's equation, we obtain:

$$\begin{aligned} \mathbb{E}[V_n] &= \mathbb{E}[F_n^+] \cdot \sigma \ , & (165)\\ \mathbb{E}[N(T_n, T_n + V_n]] &= \mathbb{E}[F_n^+] \cdot \lambda \cdot \sigma \ . & (166) \end{aligned}$$

Notice that the expectations in (165) and (166) are finite, since the visit duration is always bounded by the exponential timer.

Let $F^{*+}$ ($F^{*-}$) be the number of customers that are taken into service (served) during a visit if there are infinitely many customers waiting in the queue, and $V^*$ the duration of such a visit, i.e.,

$$\begin{aligned} 0 < \ & \lim_{x\to\infty} \mathbb{E}[f^+(x)] = \mathbb{E}[F^{*+}] \ < \infty \ , & (167)\\ 0 < \ & \lim_{x\to\infty} \mathbb{E}[f^-(x)] = \mathbb{E}[F^{*-}] \ < \infty \ , & (168) \end{aligned}$$

and also,

$$\lim_{x\to\infty} \mathbb{E}[v(x)] = \mathbb{E}[V^*] = \mathbb{E}[F^{*+}] \cdot \sigma < \infty \ . \tag{169}$$

Next, we present a lemma which will be needed in the final part of the proof. This lemma substitutes in fact Lemma 1 of [15].

**Lemma 12.** *Let $(Q_n)_n$ be a sequence of random variables converging in distribution to an (possible degenerate) integer-valued random variable $Q$. Let $(f^+, f^-, v, \phi)$ be induced by the ETL service discipline and be independent of $(Q, (Q_n)_n)$. The sequence $(Q_n, f^+(Q_n), f^-(Q_n), v(Q_n), \phi(Q_n))_n$ converges in distribution to $(Q, f^+(Q), f^-(Q), v(Q), \phi(Q))$, and*
*(i) when $\mathbb{E}[F^{*-}] < \infty$, and if $Q$ has a defective distribution, so is the limiting distribution of $Q_n - f^-(Q_n)$*
*(ii) when $\mathbb{E}[F^{*-}] < \infty$, $\mathbb{E}[F^-(Q)] < \mathbb{E}[F^{*-}]$ if and only if there exists a $y < \infty$ such that $\mathbb{P}(Q \leq y) > 0$ and $\mathbb{E}[f^-(y)] < \mathbb{E}[F^{*-}]$.*
*In both cases, if $(Q_n)_n$ is $\leq_d$-monotone, $\lim_{n\to\infty} \mathbb{E}[F^-(Q_n)] = \mathbb{E}[f^-(Q)]$ and $\lim_{n\to\infty} \mathbb{E}[v(Q_n)] = \mathbb{E}[v(Q)]$.*

*Proof.* The proof is immediate from the proof of Lemma 1 in [15]. □

**Remark 6.** *We have defined Lemma 12 in terms of the number of customers served. Analogously, this lemma can also be defined for the number of customers taken into service.*

Finally, the following lemma is essential for the remainder of the proof.

**Lemma 13.** *The ETL discipline satisfies the stochastic monotonicity property.*

*Proof.* The proof follows by sample-path arguments and is immediate from the proof of Lemma 2 in [15]. □

### D.2 Monotonicity

The stochastic monotonicity property plays in key role in the stability proof. Therefore, we will state several monotonicity results [15] which are valid for service disciplines satisfying this property.

To this end, we describe the system by the queue lengths at the polling instants and define $M(t)$ as follows:

$$M(t) = (Q_1(t), \dots, Q_M(t)),\ t \geq 0\ . \tag{170}$$

Recall that $t(i)$ denotes the queue served at visit $i$ of a cycle. We denote by visit $(n, i)$ the $i$th visit in the $n$th cycle and let visit $(1, 1)$ start at time $t = 0$. Let $T_{n,i}$ denote the time of the polling instant of visit $(n, i)$, so that we have:

$$0 = T_{1,1} \leq T_{1,2} \leq \dots \leq T_{1,a} \leq T_{2,1} \leq \dots\ . \tag{171}$$

For convenience, we write $M_{n,i}$ for $M(T_{n,i})$ and $Q_{n,i}$ for $Q_{t(i)}(T_{n,i})$. Hence, we can describe the Markovian behavior of the system as follows.

**Proposition 3.** *(Prop. 1 of [15]) The sequence $(M_{n,i})_{n,i}$ is a Markov chain. For each $i$ fixed in $\{1, \dots, a\}$, the Markov chain $(M_{n,i})_n$ is a homogeneous, aperiodic and irreducible on (a subset of) $\mathbb{N}^M$.*

*Proof.* See [15]. □

Let us define by $\pi_i$ the transition operator at visit $i$, $1 \leq i \leq a$, of the Markov chain $(M_{n,i})_{n,i}$ as follows:

$$\pi_i h(\mathbf{m}) = \mathbb{E}[h(M_{n,i+1}) | M_{n,i} = \mathbf{m}]\ , \tag{172}$$

for any $\mathbf{m} = (m_1, \dots, m_M)$ and any real function $h$ defined on $\mathbb{N}^M$ for which the expectation exists. Besides, we let $\tilde{\pi}$ be the transition operator of the Markov chain $(M_{n,i})_n$. An operator $\pi$ is said to be $\leq_d$-monotone if for all distributions $P_1 \leq_d P_2$, $\pi P_1 \leq_d \pi P_2$. This holds if $\pi h$ is $\leq$-monotone when $h$ is.

**Lemma 14.** *(Lemma 3 of [15]) For all $i$, $\pi_i$ and $\tilde{\pi}_i$ are $\leq_d$-monotone.*

*Proof.* See [15]. □

Let us next define the following quantities:

- $F^+_{n,i}$: the number of customers taken into service during visit $(n,i)$;
- $F^-_{n,i}$: the number of customers served during visit $(n,i)$;
- $V_{n,i}$: the duration of visit $(n,i)$ .

An immediate consequence of Lemma 14 is the monotonicity property of the state process.

**Proposition 4.** *Suppose $M_{1,1} = (0,\dots,0)$. Then, for each $i$, $(M_{n,i})_n$ and $(F^+_{n,i}, F^-_{n,i}, V_{n,i})$ are $\leq_d$-monotone.*

*Proof.* The proof is immediate from the proof of Prop. 2 in [15]. □

Next, we turn to dominance relations between polling systems. In particular, we compare systems with a different number of saturated queues. Here, saturation means that at a polling instant of a queue there is an infinite number of customers waiting. The saturation of a queue implies that the server serves the queue up to the time limit and then leaves. From the viewpoint of the other queues in the system, such a visit to a saturated queue is merely an additional switch-over time. Let $\mathcal{S}$ be the initial polling system with queues $1,\dots,M$. For $e \in \{0,\dots,M\}$, we define the subsystem $\mathcal{S}^e$ as the polling system consisting of the queues $1,\dots,e$, resulting from the saturation of the queues $e+1,\dots,M$, and served according to the same periodic schedule as the original system. We emphasize that if $t(i) > e$ then no queue is served but the server becomes unavailable for a duration of $V^*_{t(i)}$, which is defined as the stationary duration of a visit to queue $t(i)$ with an infinite number of customers waiting at the start of the visit. Hence, the mean total switch-over time in the subsystem $c^e_T$ can be written as:

$$c^e_T = c_T + \sum_{j=e+1}^{M} \sigma_j \mathbb{E}[G^{*+}_j] \ , \tag{173}$$

where $\mathbb{E}[G^{*+}_j]$ is the maximum expected number of customers taking into service at $Q_j$ during a cycle.

The state space of the subsystem $\mathcal{S}^e$ is given by the sequence $M^e_{n,i} = (Q^e_1(T^e_{n,i}),\dots,Q^e_e(T^e_{n,i}))$ at the polling instants $T^e_{n,i}$. For each visit $i$, $(M^e_{n,i})_n$ is a Markov chain and is $\leq_d$-monotone if the initial state is the empty state. The subsystem $\mathcal{S}^e$ is similar to the original system $\mathcal{S}$ in the sense that all previous results apply to it. Let denote by $M^{g|e}$ the $e$ first components of a vector $M^g$ having $g > e$ components. Then, the subsystems $\mathcal{S}^e$ satisfy the following dominance property.

**Lemma 15.** *(Lemma 4 of [15]) For $e < g$ both in $\{0,\dots,M\}$, $\mathcal{S}^e$ dominates $\mathcal{S}^g$ in the sense that if $M^{g|e}_{1,1} \leq_d M^e_{1,1}$ then $M^{g|e}_{n,i} \leq_d M^e_{n,i}$ for all $(n,i)$.*

*Proof.* See [15]. □

## D.3 Stability

The polling system is said to be stable if:

- (i) there exists a proper stationary joint-distribution for the queue lengths at the polling instants;
- (ii) the stationary cycle time is finite .

### D.3.1 Proof: Sufficient condition

We assume w.l.o.g. that the system is empty at time 0 as the stationary distribution of the Markov chain does not depend on the initial distribution. For convenience, let us introduce several definitions for the number of customers at a specific queue, i.e.,

- $H_k^-$ : number of customers actually served at $Q_k$ during a visit to $Q_k$;
- $H_k^+$ : number of customers taken into service at $Q_k$ during a visit to $Q_k$;
- $H_k^{*-}$ : number of customers actually served at $Q_k$ during a visit when $Q_k$ is saturated;
- $H_k^{*+}$ : number of customers taken into service at $Q_k$ during a visit when $Q_k$ is saturated .

Notice that these definitions resemble the definitions of $F_n^-, F_n^+, F_n^{*-}$ and $F_n^{*+}$. These latter quantities refer to the number of customers at the $n$th visit rather than to the number at a specific queue.

W.l.o.g. we consider the cycle from $T_{n,1}$ to $T_{n+1,1}$. Then, we may similarly define the counterparts $G_k^-$, $G_k^+$, $G_k^{*-}$ and $G_k^{*+}$ which count the same quantities but over a complete cycle. Hence, we may then also write:

$$\begin{aligned}
\mathbb{E}[G_k^-] &:= \mathbb{E}[H_{k,1}^-] + \cdots + \mathbb{E}[H_{k,a_k}^-] \ , && (174)\\
\mathbb{E}[G_k^+] &:= \mathbb{E}[H_{k,1}^+] + \cdots + \mathbb{E}[H_{k,a_k}^+] \ , && (175)\\
\mathbb{E}[G_k^{*-}] &:= \mathbb{E}[H_{k,1}^{*-}] + \cdots + \mathbb{E}[H_{k,a_k}^{*-}] \ , && (176)\\
\mathbb{E}[G_k^{*+}] &:= \mathbb{E}[H_{k,1}^{*+}] + \cdots + \mathbb{E}[H_{k,a_k}^{*+}] \ , && (177)
\end{aligned}$$

where $\mathbb{E}[H_{k,i}^-]$ is the mean number of customers served at $Q_k$ during the $i$th visit to $Q_i$ in a cycle, and $\mathbb{E}[H_{k,i}^+]$, $\mathbb{E}[H_{k,i}^{*-}]$, and $\mathbb{E}[H_{k,i}^{*+}]$ are defined similarly. Besides, we state two definitions related to the service time:

- $\sigma_k$ : mean duration of a service attempt at $Q_k$;
- $\tilde{\sigma}_k$ : mean effective service time of a customer at $Q_k$ .

A service attempt can either lead to a successful completion or to an interruption, so that:

$$\sigma_k = \mathbb{E}[min(S_k, T_k)] \ , \tag{178}$$

where $S_k$ is the original service time of a customer at $Q_k$ and $T_k$ the duration of the (exponential) timer at $Q_k$. It is good to notice that if no timer were present, then $\sigma_k = \mathbb{E}[S_k]$. The effective service time is defined as the total time spent by the server on serving a customer (including interrupted service attempts) and is in fact a geometric sum of service attempt durations. Thus, we may write for its mean:

$$\tilde{\sigma}_k = \mathbb{E}\left[\sum_{n=1}^{N} \sigma_k\right] = \sigma_k / \tilde{S}_k(\xi_k) \ , \tag{179}$$

where $N$ is geometrically distributed with success probability $p = \tilde{S}_k(\xi_k)$.

Let us denote by $\mathbb{E}[V_k^c]$ the mean total visit time to $Q_k$ during a cycle, i.e.,

$$\mathbb{E}[V_k^c] = \mathbb{E}[V_{k,1}] + \cdots + \mathbb{E}[V_{k,a_k}] \ , \tag{180}$$

where $\mathbb{E}[V_{k,j}]$ stands for the mean visit time during the $j$th visit to $Q_k$ of a cycle. Finally, we need the following mild assumption for reasons of analytical tractability:

**Assumption 1.** *Each time the server visits $Q_k$, $k = 1, \ldots, M$, the exponential time-limit is identically distributed with the same mean $1/\xi_k$.*

We note that Assumption 1 guarantees that the effective service times are identical for all customers at $Q_k$. Then, we are ready to present the following lemma:

**Lemma 16.**

$$\mathbb{E}[V_k^c] = \mathbb{E}[G_k^-] \cdot \tilde{\sigma}_k, \ k = 1, \ldots, M \ . \tag{181}$$

The proof of the lemma will be given below. However, we will derive several intermediate results first.

Clearly, when $Q_k$ is saturated, there is always exactly one interrupted service. Thus, we have the following property:

**Property 1.**

$$H_k^{*+} = H_k^{*-} + 1, \ k = 1, \ldots, M \ , \tag{182}$$

*and thus in particular:*

$$\mathbb{E}[H_k^{*+}] = \mathbb{E}[H_k^{*-}] + 1, \ k = 1, \ldots, M \ . \tag{183}$$

Besides, there is a less obvious relation between the quantities $H_k^+$, $H_k^-$, $H_k^{*+}$ and $H_k^{*-}$. However, before we get to this relation, we give a lemma and present some useful properties for $H_k^+$ and $H_k^-$.

**Lemma 17.** *Let $H$ be a geometrically distributed r.v. and let $W$ be a non-negative discrete r.v. independent $H$. Then, the following assertion holds:*

$$\mathbb{E}[H \boldsymbol{1}_{\{W \geq H\}}] + \mathbb{E}[W \boldsymbol{1}_{\{W < H\}}] = \mathbb{E}[H] \cdot \mathbb{E}[\boldsymbol{1}_{\{W \geq H\}}] \ . \tag{184}$$

*Proof.*

$$\begin{aligned} \mathbb{E}[H] &= \mathbb{E}[H\mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[H\mathbf{1}_{\{W < H\}}] && (185) \\ &= \mathbb{E}[H\mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[W\mathbf{1}_{\{W < H\}}] + \mathbb{E}[(H - W)\mathbf{1}_{\{W < H\}}] \ . && (186) \end{aligned}$$

Next, we may use the fact that $H$ is a geometric and thus memoryless random variable, i.e., $H - W|_{H > W} =_d H$, so that:

$$\mathbb{E}[H] = \mathbb{E}[H\mathbf{1}_{\{W \geq H\}}] + \mathbb{E}[W\mathbf{1}_{\{W < H\}}] + \mathbb{E}[H] \cdot \mathbb{E}[\mathbf{1}_{\{W < H\}}] \ . \tag{187}$$

This completes the proof. □

Denote by $V$ the number of customers served until $Q_k$ would become empty for the first time if there were no timer. Then, the following properties are readily verified:

**Property 2.**

$$H_k^+ = \min(V, H_k^{*+}) = V\boldsymbol{1}_{\{V \leq H_k^{*+}\}} + H_k^{*+}\boldsymbol{1}_{\{V > H_k^{*+}\}} \ , \tag{188}$$

$$H_k^- = \min(V, H_k^{*-}) = V\boldsymbol{1}_{\{V \leq H_k^{*-}\}} + H_k^{*-}\boldsymbol{1}_{\{V > H_k^{*-}\}} \ . \tag{189}$$

These properties imply that if the server leaves $Q_k$ because it is empty, then $H_k^+ = H_k^-(= V)$ and $H_k^+ = H_k^- + 1$, otherwise.

The following lemma demonstrates that the ratio of mean number of served customers and mean number of customers taken into service is equal both for a saturated and a non-saturated queue.

**Lemma 18.**

$$\frac{\mathbb{E}[H_k^{*-}]}{\mathbb{E}[H_k^{*+}]} = \frac{\mathbb{E}[H_k^-]}{\mathbb{E}[H_k^+]}, \ k = 0, 1, \ldots, M \ . \tag{190}$$

*Proof.* Note that $H_k^{*+}$ is a geometrically distributed random variable (with success probability $p = 1 - \tilde{S}_k(\xi_k)$, since an interruption is seen as a success). Then,

$$\begin{aligned}
\mathbb{E}[H_k^+] \cdot \mathbb{E}[H_k^{*-}] &= \Big(\mathbb{E}[V\mathbf{1}_{\{V<H_k^{*+}\}}] + \mathbb{E}[H_k^{*+}\mathbf{1}_{\{V\geq H_k^{*+}\}}]\Big)\left(\mathbb{E}[H_k^{*+}] - 1\right) && (191)\\
&= \Big(\mathbb{E}[V\mathbf{1}_{\{V<H_k^{*+}\}}] + \mathbb{E}[H_k^{*+}\mathbf{1}_{\{V\geq H_k^{*+}\}}]\Big) \cdot \mathbb{E}[H_k^{*+}] && (192)\\
&\quad - \Big(\mathbb{E}[V\mathbf{1}_{\{V<H_k^{*+}\}}] + \mathbb{E}[H_k^{*+}\mathbf{1}_{\{V\geq H_k^{*+}\}}]\Big) && (193)\\
&= \Big(\mathbb{E}[V\mathbf{1}_{\{V<H_k^{*+}\}}] + \mathbb{E}[H_k^{*+}\mathbf{1}_{\{V\geq H_k^{*+}\}}]\Big) \cdot \mathbb{E}[H_k^{*+}] - \mathbb{E}[\mathbf{1}_{\{V\geq H_k^{*+}\}}] \cdot \mathbb{E}[H_k^{*+}] && (194)\\
&= \mathbb{E}[H_k^{*+}] \cdot \Big(\mathbb{E}[V\mathbf{1}_{\{V<H_k^{*+}\}}] + \mathbb{E}[(H_k^{*+} - 1)\mathbf{1}_{\{V\geq H_k^{*+}\}}]\Big) \ , && (195)
\end{aligned}$$

where for the third equality sign we used Lemma 17. Finally, observe that $\{V < H_k^{*+}\} \Leftrightarrow \{V \leq H_k^{*-}\}$ (since all variables are discrete), so that we may write:

$$\begin{aligned}
&\mathbb{E}[V\mathbf{1}_{\{V<H_k^{*+}\}}] + \mathbb{E}[(H_k^{*+} - 1)\mathbf{1}_{\{V\geq H_k^{*+}\}}] && (196)\\
&= \mathbb{E}[V\mathbf{1}_{\{V\leq H_k^{*-}\}}] + \mathbb{E}[(H_k^{*+} - 1)\mathbf{1}_{\{V> H_k^{*-}\}}] && (197)\\
&= \mathbb{E}[V\mathbf{1}_{\{V\leq H_k^{*-}\}}] + \mathbb{E}[H_k^{*-}\mathbf{1}_{\{V> H_k^{*-}\}}] && (198)\\
&= \mathbb{E}[H_k^-] \ , && (199)
\end{aligned}$$

where we used Prop. 2 in the final step. □

**Remark 7** (Independence of $V$)**.** *It is important to notice that the equivalence of the ratios is independent of $V$. In particular, we have made no assumptions whatsoever on the number of customers present at the start of a visit in the unsaturated case. So, the time of the previous polling instant of the queue does not impact the ratio of the unsaturated case (while the ratio of the saturated case is obviously fixed).*

Recall that $\mathbb{E}[V_k^*]$ denotes the mean visit time of the server to $Q_k$ when $Q_k$ is saturated. This quantity satisfies the following relation.

**Lemma 19.**

$$\mathbb{E}[V_k^*] = \mathbb{E}[H_k^{*+}] \cdot \sigma_k \;=\; \mathbb{E}[H_k^{*-}] \cdot \tilde{\sigma}_k,\ k = 1, \ldots, M\ . \tag{200}$$

*Proof.* First, consider the saturated case. We write: $V_k^* = \sum_{j=1}^{H_k^{*+}} S_{k,j}$, where $S_{k,j},\ j = 1, 2, \ldots,$ is iid r.v. distributed as $min(S_k, T_k)$. Observe that $H_k^{*+}$ is a stopping time for $S_{k,j},\ j = 1, 2, \ldots$, so that we may apply Wald's equation yielding: $\mathbb{E}[V_k^*] = \mathbb{E}[H_k^{*+}] \cdot \sigma_k$. Next, consider a period $T$ comprising a single visit of length $V_k$ to $Q_k$ extended with the (service) time needed to complete the service of the customer that was interrupted at the end of the visit. That is, $T$ is the time needed to complete all services that were (re-)started during $V$ (in particular, we include a possible residual service time, which is in fact distributed as an original service time due to the geometric nature of the effective service time). Thus, $T = \mathbb{E}[H_k^{*+}] \cdot \tilde{\sigma}_k$, but also $T = \mathbb{E}[V_k] + \tilde{\sigma}_k$, since there is always an interrupted service with a mean residual service time identical to the original mean effective service time. Hence, it follows that: $\mathbb{E}[V_k] = (\mathbb{E}[H_k^{*+}] - 1) \cdot \tilde{\sigma}_k = \mathbb{E}[H_k^{*-}] \cdot \tilde{\sigma}_k$. □

*Proof.* (Proof of Lemma 16) It is readily seen that to prove Eq. (181) it is sufficient to show:

$$\mathbb{E}[V_{k,j}] = \mathbb{E}[H_{k,j}^-] \cdot \tilde{\sigma}_k,\ j = 1, \ldots, a_k\ . \tag{201}$$

W.l.o.g. we consider the first visit to $Q_k$ in a cycle and leave out the subscript 1. Thus, we need to prove the following:

$$\mathbb{E}[V_k] = \mathbb{E}[H_k^-] \cdot \tilde{\sigma}_k\ . \tag{202}$$

Analogously to the proof of Lemma 19, we write $V_k = \sum_{j=1}^{H_k^+} S_{k,j}$ for the unsaturated case. By arguing that $H_k^+$ is a stopping time for the sequence $\{S_{k,j}\}_j$, it immediately follows via Wald that: $\mathbb{E}[V_k] = \mathbb{E}[H_k^+] \cdot \sigma_k$. The proof is then completed by appealing to Lemma 18 and Lemma 19. □

Let us define $\hat{\rho}_k,\ k = 1, \ldots, M$ as follows:

$$\hat{\rho}_k := \sum_{j=1}^{k} \rho_j = \sum_{j=1}^{k} \lambda_j \tilde{\sigma}_j\ . \tag{203}$$

Next, we introduce a stability condition for the complete system:

**Definition 3.** *(Condition $\mathcal{C}^M$)*

$$\mathcal{C}^M : \hat{\rho}_M + \max_{1 \le j \le M} (\lambda_j / \mathbb{E}[G_j^{*-}]) c_T < 1\ , \tag{204}$$

and one for the subsystem $\mathcal{S}^e$ with $e \in \{0, \ldots, M\}$:

**Definition 4.** *(Condition $\mathcal{C}^e$)*

$$\mathcal{C}^e : \hat{\rho}_e + \max_{1 \le j \le e} (\lambda_j / \mathbb{E}[G_j^{*-}]) c_T^e < 1 \ . \tag{205}$$

We number the queues according to the ratio $\lambda_j / \mathbb{E}[G_j^{*-}]$ in non-decreasing order. Hence, we have that:

$$\mathcal{C}^e : \hat{\rho}_e + (\lambda_e / \mathbb{E}[G_e^{*-}]) c_T^e < 1 \ , \tag{206}$$

Further, we note that it can verified by simple calculations that $\mathcal{C}^{e+1}$ implies $\mathcal{C}^e$

We are now ready to present the following lemma (cf. Lemma 6 of [15]) which forms a crucial link in the proof:

**Lemma 20.** *If condition $\mathcal{C}^e$ holds, then*

$$\mathbb{E}[G_k^{e-}] < \mathbb{E}[G_k^{*-}], \ 1 \le k \le e \ . \tag{207}$$

*Proof.* Let us consider the mean duration of a cycle. W.l.o.g. we say that the $n$th cycle starts at time $T_{n,1}$ and ends at time $T_{n+1,1}$. A cycle consists of the visits to the queues and the switch-over times, so that we may write (cf. Lemma 16):

$$\mathbb{E}[T_{n+1,1} - T_{n,1}] = \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T, \ n = 0, 1, \dots \ . \tag{208}$$

Hence, for the change in number of customers at $Q_k$ during this cycle, we readily have:

$$\mathbb{E}[Q_k(T_{n+1,1}) - Q_k(T_{n,1})] = \lambda_k \cdot \left( \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T \right) - \mathbb{E}[G_{n,k}^-], \ k = 1, \dots, M, \ n = 0, 1, \dots \ . \tag{209}$$

Suppose w.l.o.g. that the system is empty at time 0, then using the $\le_d$-monotonicity for each given visit, it follows that the expectations of the queue lengths at the polling times are non-decreasing, i.e.,

$$\mathbb{E}[Q_k(T_{n+1,1}) - Q_k(T_{n,1})] \ge 0, \ k = 1, \dots, M, \ n = 0, 1, \dots \ . \tag{210}$$

which provides us immediately with the following system of equations:

$$\mathbb{E}[G_{n,k}^-] \le \lambda_k \cdot \left( \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T \right), \ k = 1, \dots, M, \ n = 0, 1, \dots \ . \tag{211}$$

Observe that $\mathbb{E}[G_{n,k}^-]$ and $\mathbb{E}[G_{n,k}^+]$ are non-decreasing in $n$ and are bounded from above by $\mathbb{E}[G_k^{*+}] < \infty$. Thus, we may define the following limits for $k = 1, \dots, M$:

$$\mathbb{E}[G_k^-] \ = \ \lim_{n \to \infty} \mathbb{E}[G_{n,k}^-] \tag{212}$$

$$\mathbb{E}[G_k^+] \ = \ \lim_{n \to \infty} \mathbb{E}[G_{n,k}^+] \ . \tag{213}$$

Let us consider next Eq. (211) for $k = 1$ and $n \to \infty$:

$$\mathbb{E}[G_1^-] \le \lambda_1 \cdot \left( \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right) \ . \tag{214}$$

This can readily be rewritten to:

$$\mathbb{E}[G_1^-] \cdot (1 - \hat{\rho}_1) \le \lambda_1 \cdot \left( \sum_{j=2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right) \ . \tag{215}$$

Applying a triangularization procedure (see Appendix D.4), we may obtain for $1 \le k \le M$:

$$\mathbb{E}[G_k^-] \cdot (1 - \hat{\rho}_k) \le \lambda_k \cdot \left( \sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right) \ . \tag{216}$$

The latter result does also hold if we consider the system with queues $e + 1$ up to $M$ being saturated. From the point of view of the first $e$ queues only the return time of the server will change while the behavior of the server during a visit remains identical. Denoting the quantities in this modified system by adding the superscript $e$, we may write:

$$\mathbb{E}[G_k^{e-}] \le \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot \left( \sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^{e-}] + c_T^e \right), \ k = 1, \ldots, e \ , \tag{217}$$

where $c_T^e = c_T + \sum_{j=e+1}^{M} \mathbb{E}[V_j^*]$. Since, $\mathbb{E}[G_j^{e-}] \le \mathbb{E}[G_j^{*-}], \ j = 1, \ldots, M$, we obtain:

$$\begin{aligned} \mathbb{E}[G_k^{e-}] \ &\le \ \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot \left( \sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^{*-}] + c_T \right) && (218) \\ &= \ \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot c_T^k, \ k = 1, \ldots, e \ . && (219) \end{aligned}$$

On the other hand, the condition $\mathcal{C}^k, \ k = 1, \ldots, e$, which is implied by $\mathcal{C}^e$, reads:

$$\mathcal{C}^k : \hat{\rho}_k + \max_{1 \le j \le k} (\lambda_j / \mathbb{E}[G_j^{*-}]) \cdot c_T^k < 1 \ . \tag{220}$$

Under the assumption that the ratios $\lambda_j / \mathbb{E}[G_j^{*-}]$ are ordered non-decreasingly, it is readily found that $\mathcal{C}^k$ implies:

$$\mathbb{E}[G_k^{*-}] > \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot c_T^k \ , \tag{221}$$

which completes the proof. □

**Remark 8.** *Notice that the key element in the proof of Lemma 20 is to obtain a strict inequality; the inequality is obvious.*

**Rest of sufficiency proof** Recall that we want to show here that if condition $\mathcal{C}^M$ is satisfied, then the system is stable. An equivalent definition of stability (see [15]) is that there exists a proper stationary joint queue-length distribution at the polling instants such that the expectation of the stationary cycle time is finite. A sufficient condition for the stationary distribution to exist is that the multi-dimensional Markov chain $(M^e_{n,1})$ is ergodic. The ergodicity of this chain $(M^e_{n,1})$ is equivalent to the existence of $\mathbf{m}^e$ such that the limit

$$\lim_{n\to\infty} \mathbb{P}(M^e_{n,1} \leq \mathbf{m}^e) \geq 1 - \sum_{k=1}^{e} \lim_{n\to\infty} \mathbb{P}(Q_k(T^e_{n,1}) \geq \mathbf{m}_k) \ , \tag{222}$$

is strictly positive. We note that the system will become empty once the chain enters some state $\leq \mathbf{m}^e$ with a strictly positive probability (due to no arrivals for a specific time). Since the positiveness of the limit implies that you return infinitely often to some state $\leq \mathbf{m}^e$, it follows that with probability one you will reach the empty state in a finite amount of time; in other words, it excludes transient or null-recurrent behaviour of the chain. Thus, to have ergodicity, we need the sum on the right-hand side to be strictly smaller than one. This can be established if for one $k$ the limiting distribution $(Q_k(T^e_{n,1}))_n$ is not concentrated at infinity, i.e, $\mathbb{P}(Q_k < \infty) > 0$ and all other limiting distributions are proper, i.e., $\mathbb{P}(Q_k < \infty) = 1$.

We will prove this by induction starting with the subsystem $\mathcal{S}^0$. This system $\mathcal{S}^0$ is readily seen to be stable. Next, we suppose $\mathcal{S}^{e-1}$ is stable, and consider $\mathcal{S}^e, \ 1 \leq e \leq M$. We note since $\mathcal{S}^{e-1}$ is stable, the Markov chain $(M^{e-1}_{n,1})$ is ergodic and in particular $(Q_k(T^{e-1}_{n,1}))_n, \ 1 \leq k \leq e-1$ has a proper distribution. Also, $(M^{e-1}_{n,i})_n, \ i = 1, ..., a$ has a proper limiting distribution and by Lemma 15, $M^{e|e-1}_{n,i} \leq_d M^{e-1}_{n,i}$ for all $n$. Thus, $(M^{e|e-1}_{n,i})_n$ has a proper limiting distribution. Moreover, from Lemma 16, we have that $\mathbb{E}[G^{e-}_e] < \mathbb{E}[G^{*-}_e]$. Hence, there exists a visit $r$ such that $\lim_{n\to\infty} \mathbb{E}[F^{e-}_{n,r}] < \mathbb{E}[F^{*-}_r]$. Then, by Lemma 12-ii there exists a $y$ such that $\lim_{n\to\infty} \mathbb{P}(Q^e_{n,r} \leq y) > 0$, i.e., the limiting distribution of the last component $Q^e_{n,r} = Q^e_e(T_{n,r})$ of $M^e_{n,r}$ is not concentrated at infinity. Thus, the chain $(M^e_{n,r})_n$ is ergodic. The observation that the expectation of the cycle time is finite completes the proof.

### D.3.2 Proof: Necessary condition

Suppose the polling system $\mathcal{S}$ is stable. Let us define $F^-_{n,k_l}$ as the mean number of customers served during the $k_l$-th stage of the $n$th cycle, where the $k_l$-th stage corresponds to exactly the $l$th visit to $Q_k$ in the cycle. We let for each visit $i$ the initial distribution of $(M_{n,i})_n$ be its stationary distribution. Since $\mathcal{S}$ is stable, these chains are stationary with positive-recurrent states. As a result, $\mathbb{P}(Q_k(T_{n,i}) = 0) > 0$ for all $k$ and $(n,i)$. Further, as the expected cycle time is finite, $\mathbb{E}[G^-_k] = \sum_{l=1}^{a_k} \mathbb{E}[F^-_{n,k_l}]$ does not depend on $n$ and is finite for all $k$. It follows by Lemma 12 that $\mathbb{E}[G^-_k] < \mathbb{E}[G^{*-}_k]$ for $1 \leq k \geq M$ and in particular that $\mathbb{E}[G^-_M] < \mathbb{E}[G^{*-}_M]$.

On the other hand, it can be readily be seen that:

$$Q_k(T_{2,1}) - Q_k(T_{1,1}) = N_k(T_{1,1}, T_{2,1}] - \sum_{l=1}^{a_k} F^-_{1,k_l} \ . \tag{223}$$

Hence, we can bound $Q_k(T_{2,1}) - Q_k(T_{1,1})$ as follows:

$$-\sum_{l=1}^{a_k} F^-_{1,k_l} \leq Q_k(T_{2,1}) - Q_k(T_{1,1}) \leq N_k(T_{1,1}, T_{2,1}] \ . \tag{224}$$

Both the lower and upper bound have finite expectation, such that for all $k$ (see [15, Lemma 7]):

$$Q_k(T_{2,1}) - Q_k(T_{1,1}) = 0 \ , \tag{225}$$

and in general for $n \geq 1$:

$$Q_k(T_{n+1,1}) - Q_k(T_{n,1}) = 0 \ . \tag{226}$$

This leads to (cf. Eq. (211)) the following system of equalities:

$$\mathbb{E}[G_k^-] = \lambda_k \cdot \left( \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right) , \ 1 \leq k \leq M \ , \tag{227}$$

and along the lines of deriving Eq. (216), we obtain:

$$\mathbb{E}[G_k^-] \cdot (1 - \hat{\rho}_k) = \lambda_k \cdot \left( \sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right) , \ 1 \leq k \leq M \ . \tag{228}$$

Specifically, for $k = M$ this implies:

$$\mathbb{E}[G_M^-] \cdot (1 - \hat{\rho}_M) = \lambda_M \cdot S \ . \tag{229}$$

Together with observation above, $\mathbb{E}[G_M^-] < \mathbb{E}[G_M^{*-}]$, it follows that condition $\mathcal{C}^M$ holds.

## D.4 Triangularization

Let us explain below the triangularization method that we apply. We depart from the following set of equalities:

$$\mathbb{E}[G_k^-] \leq \lambda_k \cdot \left( \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) , \ k = 1, \ldots, M \ . \tag{230}$$

Rearranging the equality for $k = 1$, we obtain:

$$(1 - \hat{\rho}_1) \cdot \mathbb{E}[G_1^-] \leq \lambda_1 \cdot \left( \sum_{j=2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \ . \tag{231}$$

Next, we will show that also for $2 \leq k \leq M$ we may write:

$$(1 - \hat{\rho}_k) \cdot \mathbb{E}[G_k^-] \leq \lambda_k \cdot \left( \sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \ . \tag{232}$$

This will be done by proving the following inequalities by induction.

$$\sum_{j=1}^{k} \tilde{\sigma}_j \mathbb{E}[G_j^-] \ \leq \ \frac{\hat{\rho}_k}{1 - \hat{\rho}_k} \cdot \left( \sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) , \ k = 1, \ldots, M \ , \tag{233}$$

$$(1 - \hat{\rho}_{k+1}) \cdot \mathbb{E}[G_{k+1}^-] \ \leq \ \lambda_{k+1} \cdot \left( \sum_{j=k+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) , \ k = 1, \ldots, M - 1 \ . \tag{234}$$

First, notice that for $k = 1$ Eq. (233) has been shown above, while Eq. (234) reads as follows.

$$(1 - \hat{\rho}_2) \cdot \mathbb{E}[G_2^-] \leq \lambda_2 \cdot \left( \sum_{j=3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) . \tag{235}$$

This inequality can be proven from Eq. (230) and taking $k = 2$. First, we take all terms $\mathbb{E}[G_2^-]$ to the left-hand side, second we apply Eq. (231), and finally some simple manipulations provide us with the desired result. Next, we show that once these inequalities hold for $l$ these also hold for $l + 1$. First, consider Eq. (233) for $l + 1$:

$$\begin{aligned}
\sum_{j=1}^{l+1} \tilde{\sigma}_j \mathbb{E}[G_j^-] &= \sum_{j=1}^{l} \tilde{\sigma}_j \mathbb{E}[G_j^-] + \tilde{\sigma}_{l+1} \mathbb{E}[G_{l+1}^-] & (236)\\
&\leq \frac{\hat{\rho}_l}{1 - \hat{\rho}_l} \cdot \left( \sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \frac{1}{1 - \hat{\rho}_l} \cdot \tilde{\sigma}_{l+1} \mathbb{E}[G_{l+1}^-] & (237)\\
&\leq \left( \frac{\hat{\rho}_l}{1 - \hat{\rho}_l} + \frac{\rho_{l+1}}{(1 - \hat{\rho}_l)(1 - \hat{\rho}_{l+1})} \right) \cdot \left( \sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) & (238)\\
&= \frac{\hat{\rho}_{l+1}}{1 - \hat{\rho}_{l+1}} \cdot \left( \sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) . & (239)
\end{aligned}$$

Second, we have to prove Eq. (234) for $l + 1$, i.e.,

$$(1 - \hat{\rho}_{l+2}) \cdot \mathbb{E}[G_{l+2}^-] \leq \lambda_{l+2} \cdot \left( \sum_{j=l+3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) . \tag{240}$$

To this end, we depart from Eq. (230) for $l + 2$:

$$\begin{aligned}
\mathbb{E}[G_{l+2}^-] &\leq \lambda_{l+2} \cdot \left( \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) & (241)\\
&= \lambda_{l+2} \cdot \sum_{j=1}^{l+1} \tilde{\sigma}_j \mathbb{E}[G_j^-] + \lambda_{l+2} \cdot \left( \sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) & (242)\\
&\leq \lambda_{l+2} \cdot \frac{\hat{\rho}_{l+1}}{1 - \hat{\rho}_{l+1}} \cdot \left( \sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \lambda_{l+2} \cdot \left( \sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) & (243)\\
&= \frac{1}{\hat{\rho}_{l+1}} \cdot \lambda_{l+2} \cdot \left( \sum_{j=l+3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \frac{1}{\hat{\rho}_{l+1}} \cdot \lambda_{l+2} \cdot \tilde{\sigma}_{l+2} \mathbb{E}[G_{l+2}^-] . & (244)
\end{aligned}$$

Hence, moving all the terms $\mathbb{E}[G_{l+2}^-]$ to the left-hand side and performing some rearrangements yields Eq. (240).

# References

[1] Delay Tolerant Networking Research Group, website: http://www.dtnrg.org.

[2] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren, "A tandem queueing model for delay analysis in disconnected ad hoc networks," in *Proc. of ASMTA*, Nicosia, Cyprus, 2008.

[3] H. Takagi, "Queueing analysis of polling systems: An update," *Chapter of Stochastic Analysis of Computer and Communication Systems*, pp. 267–318, 1990.

[4] ——, "Queueing analysis of polling models: progress in 1990-1994," in *Frontiers in Queueing: Models, Methods and Problems, J.H. Dshalalow (ed.).* CRC Press, Boca Raton, 1997, pp. 119–146.

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

[6] M. Eisenberg, "Queues with periodic service and changeover times," *Operations Research*, vol. 20(2), pp. 440–451, 1972.

[7] S. W. Fuhrmann, "A decomposition result for a class of polling models," *Queueing Systems*, vol. 11(1-2), pp. 109–120, 1992.

[8] E. de Souza e Silva, H. R. Gail, and R. R. Muntz, "Polling systems with server timeouts and their application to token passing networks," *IEEE Trans. on Networking*, vol. 3(5), pp. 560–575, 1995.

[9] K. K. Leung, "Cyclic-service systems with non-preemptive time-limited service," *IEEE Trans. on Communications*, vol. 42(8), pp. 2521–2524, 1994.

[10] I. Eliazar and U. Yechiali, "Randomly timed gated queueing systems," *SIAM Journal of Applied Mathematics*, vol. 59(2), pp. 423–441, 1998.

[11] ——, "Polling under the randomly timed gated regime," *Stochastic Models*, vol. 14(1-2), pp. 79–93, 1998.

[12] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren, "Time-limited and k-limited polling systems: A matrix geometric solution," in *Proc. of SMCTools*, Athens, Greece, 2008.

[13] J. W. Cohen, *The Single Server Queue*, 2nd ed. Elsevier Science Publishers, 1992.

[14] J. Resing, "Polling systems and multitype branching processes," *Queueing Systems*, vol. 13(4), pp. 409–426, 1993.

[15] C. Fricker and M. R. Jaïbi, "Monotonicity and stability of periodic polling systems," *Queueing Systems*, vol. 15(1-4), pp. 211–238, 1994.