
# Polling systems with multiple coupled servers

S.C. Borst

Department of Operations Research, Statistics, and System Theory

**Report BS-R9408 February 1994**

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.


P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

# Polling Systems with Multiple Coupled Servers

S.C. Borst
*CWI*
*P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*

**Abstract**

We consider polling systems with multiple coupled servers. We explore the class of systems that allow an exact analysis. For these systems we present distributional results for the waiting time, the marginal queue length, and the joint queue length at polling epochs. The class in question includes several single-queue systems with a varying number of servers, two-queue two-server systems with exhaustive service and exponential service times, as well as infinite-server systems with an arbitrary number of queues, exhaustive or gated service, and deterministic service times.

*AMS Subject Classification (1991):* 60K25, 68M20.
*Keywords & Phrases:* polling system, multiple servers, queue length, waiting time.

## 1 INTRODUCTION

In this paper we consider polling systems with multiple coupled servers. A multiple-server polling system is a multiple-queue system attended to by multiple servers in a cyclic manner. Like the single-server versions, multiple-server polling systems find applications in computer systems, communication networks, and manufacturing environments. So far there are hardly any exact results known for these systems, apart from some mean value results for global performance measures like cycle times. In this paper we explore the class of systems that allow an exact analysis. For these systems we present distributional results for the waiting time, the marginal queue length, and the joint queue length at polling epochs. In the remainder of the introduction we successively discuss some applications, present an overview of related literature, and describe the organization of the paper.

*Applications*

An example of a multiple-server polling system is a distributed system, consisting of a number of computers, interconnected by a communication medium, that cooperate as follows in sharing the total load of the system, cf. [23]. The jobs entering the 'front-end' systems (corresponding to the queues) are picked up in batches by the 'back-end' systems (corresponding to the servers) according to some cyclic schedule. As soon as a batch is served, the back-end system picks up the jobs from the next front-end system. When the communication delay is negligible, it may be preferable to use a more sophisticated strategy which keeps track of the evolution of the system, like "serve the longest queue". However, when the communication

Report BS-R9408
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

overhead is significant, it makes sense to use a cyclic service schedule which requires only very little information processing.
Examples also arise in communication networks, like the underlying communication medium in the above-mentioned distributed system. Consider e.g. a Local Area Network (LAN), consisting of a number of stations, interconnected by a transmission ring. There are various protocols known for the medium access control in a LAN with a ring architecture. One variant is the slotted ring, i.e., the ring is subdivided into time slots of the size of a single message, circulating at constant speed. When a station sees an empty slot pass by, it may put a message in it. In case of destination release the receiving station subsequently removes the message from the slot, whereas in case of source release the transmitting station empties the slot. So occupying a slot corresponds to utilizing a server. Thus a slotted ring may be viewed as a multiple-server polling system (unless there is only a single slot). Another medium access variant that may lead to multiple-server polling is the token ring, i.e., there is a token circulating on the ring, representing the right of transmission. So holding the token corresponds to utilizing the server. Thus a token ring is in fact a single-server system, but the stations may happen to be interconnected by multiple token rings rather than only a single token ring.
Other examples arise in manufacturing environments. In flexible manufacturing systems e.g. one finds a group of machines that periodically change over from one type of operations to another. In some factories internal transport is provided by a chain of vehicles that follow a track along loading/unloading points, which conceptually comes very close to a slotted ring with destination release. A multiple-elevator facility is yet another example, although some features, like the acceleration effect when a floor is skipped, are not incorporated in the classical description of a polling model, cf. [15].

*Related literature*
Multiple-server polling systems have received remarkably little attention in the vast literature on polling systems. One of the first studies is Morris & Wang [23] in which the servers are assumed to be independent, i.e., to visit the queues independently of each other, each server according to some cyclic schedule. They obtain the mean cycle time of each server as well as the mean intervisit time to a queue, and derive approximate expressions for the mean sojourn time for both a gated-type and a limited-type service discipline. A very interesting phenomenon observed by Morris & Wang is the tendency for the servers to cluster if they follow identical routes, especially in heavy traffic. The phenomenon may be visualized as follows. A trailing server will tend to move fast, as it only encounters recently served queues, whereas a leading server will tend to be slowed down by queues that have not been served for a while, so that the servers tend to form bunches while constantly leapfrogging over one another. Numerical experiments indicate that the bunching of servers is likely to deteriorate the system performance. Obviously the bunching of servers is alleviated if they follow different routes. Therefore Morris & Wang advocate the use of 'dispersive' schedules to improve the system performance.
Browne & Weiss [12] is one of the few studies in which the servers are assumed to be coupled, i.e., to visit the queues together. They obtain index rules for the minimization of the mean length of individual cycles for both the exhaustive and the gated service discipline. Browne et al. [10] derive the mean waiting time for a completely symmetric two-queue system with

an infinite number of coupled servers and deterministic service times. Browne & Kella [11] obtain the busy period distribution for a two-queue system with an infinite number of coupled servers, exhaustive service, and deterministic service times at one queue and general service times at the other queue.
Levy & Yechiali [19] and Kao & Narayanan [17] study the joint distribution of the queue length and the number of busy servers for a Markovian multiple-server queue, where the servers individually go on vacation when there are no waiting customers left. Mitrany & Avi-Itzhak [22] and Neuts & Lucantoni [24] analyze the joint distribution of the queue length and the number of busy servers for a Markovian multiple-server queue where servers break down at exponential intervals and then get repaired. In references [6], [16], [18], [25], [29] mean response time approximations are developed to analyze the performance of LAN's with multiple token rings. Mean response time approximations oriented to LAN's with a multiple slotted ring are contained in references [5], [6], [20], [29], [30]. Ajmone Marsan et al. [2], [3], [4] derive the mean cycle time and bounds for the mean waiting times in symmetric systems for the exhaustive, gated, and 1-limited service discipline. In [1] they illustrate how PETRI-net techniques may be used to study Markovian multiple-server polling systems.
The above-mentioned studies unanimously point out that multiple-server polling systems, combining the complexity of single-server polling systems and multiple-server systems, are extraordinarily hard to analyze. In fact almost none of the studies (except [10], [11] and the single-queue studies [17], [19], [22], [24]) presents any exact results, apart from some mean value results for global performance measures like cycle times.

*Organization of the paper*
In the present paper we consider the case of coupled servers. We are mainly interested in exploring the class of systems that allow an exact analysis. For these systems we present distributional results for the waiting time, the marginal queue length, and the joint queue length at polling epochs. The motivation for considering the case of coupled servers is three-fold. First, the dependence in the position of the servers does not play any complicating role then. Second, in some situations the servers may in fact happen to be physically coupled. Third, the coupled server case may also be relevant for the study of the independent server case. The tendency for the servers to cluster provides e.g. an indication that they tend to behave as if they were coupled.
The remainder of the paper is organized as follows. In section 2 we consider a single-queue multiple-server system with service interruptions, which is not only interesting in its own right but also useful for the study of a multiple-server polling system. In isolation a particular queue in a polling system may be viewed as a single-queue system with service interruptions, the intervisit periods constituting the service interruptions. Results for single-queue systems with service interruptions may thus be used to obtain results for the marginal distributions in polling systems. In section 3 we return to the multiple-server polling system. We relate the probability generating function (pgf) of the joint queue length at the beginning of a visit to the pgf of the joint queue length at the end of the *previous* visit. Next we relate the pgf of the joint queue length distribution at the *end* of a visit to the pgf of the joint queue length at the *beginning* of a visit. Thus we obtain $2n$ equations involving $2n$ pgf's, with $n$ the number of queues. In section 4 we identify some cases for which these pgf's can actually be solved from these equations. These cases include several single-queue systems with a varying number of

servers, two-queue two-server systems with exhaustive service and exponential service times, as well as infinite-server systems with an arbitrary number of queues, exhaustive or gated service, and deterministic service times. In section 5 we conclude with some remarks and suggestions for further research.

## 2 AN $M/M/m$ QUEUE WITH COUPLED SERVERS AND SERVICE INTERRUPTIONS

In the present section we consider an $M/M/m$ queue with coupled servers and service interruptions. The service interruptions are assumed to result from some interfering process that from time to time keeps the servers from working, even while there are customers present. Service preemption due to service interruptions is allowed. The service interruptions may be interwoven with the arrival and service processes in an arbitrarily complex manner, but may not anticipate on the future arrival and service times of customers. In particular the *durations* of successive service interruptions are allowed to be dependent. We abstract here from what kind of interfering process causes the service interruptions. In the context of polling models a service interruption typically models the intervisit period. In the setting of performability models a service interruption usually represents a down-period of the system. A period during which none of the servers is busy, because of a service interruption or because there are no customers present, will be called a non-serving interval. A period during which at least one of the servers is busy, will be called a serving interval.

Fuhrmann & Cooper [14] consider an $M/G/1$ queue with service interruptions. Under rather mild assumptions they prove a decomposition property of the queue length distribution. Using concepts from the theory of branching processes they show that the queue length distribution can be expressed as the convolution of the distribution of the following two quantities:

(i). the queue length at an arbitrary epoch in the 'corresponding' $M/G/1$ queue without service interruptions;

(ii). the queue length at an arbitrary epoch in a non-serving interval.

The 'corresponding' $M/G/1$ queue without service interruptions is an ordinary $M/G/1$ queue with similar traffic characteristics, of which the queue length distribution is simply known from the Pollaczek-Khintchine formula. To find the queue length distribution at an arbitrary epoch, it thus suffices to find the queue length distribution in a non-serving interval, which is quite often relatively simple. By the distributional form of Little's law the queue length decomposition also translates into a decomposition of the waiting time. Under somewhat milder assumptions than Fuhrmann & Cooper, Boxma [8] proves a similar decomposition of the amount of work in the system. Browne & Kella [11] analyze the queue length distribution in an $M/G/\infty$ queue with vacations. They observe that for deterministic service times a Fuhrmann & Cooper like decomposition property holds but not for exponential service times.

We now analyze the queue length distribution in the $M/M/m$ queue under consideration. Although for $m = 1$ the amount of work is somewhat easier to study than the queue length, for $m > 1$ we need to focus on the queue length, as the amount of work then no longer completely determines the number of busy servers. We make the following assumptions.

(i). During a serving interval there are no servers idling while there are customers waiting, i.e., if there are $l$ customers present during a serving interval then there are $\min(l, m)$ servers working, just like in an ordinary $M/M/m$ queue.

(ii). The order in which customers enter service is independent of their service times.

Under the above assumptions we will show that the queue length distribution can be expressed into the distribution of (conceptually) the same two quantities as in the $M/G/1$ queue with service interruptions, but not in the same simple convolution form. However, to find the queue length distribution at an arbitrary epoch, it still suffices to find the queue length distribution in a non-serving interval. Under some additional assumptions we will also show how the queue length decomposition translates into a decomposition of the waiting time.

We first introduce some notation. Let $\lambda$ be the arrival rate and let $\mu$ be the service rate. Define $\rho := \lambda/\mu$. Denote by $\mathbf{N}$ and $\mathbf{N}_A$ the total number of customers present (including customers in service) at, respectively, an arbitrary epoch and an arbitrary epoch in a non-serving interval. Denote by $\mathbf{N}^{(l)}_{M/M/m}$ the number of customers at an arbitrary epoch in the 'corresponding' $M/M/m$ queue, given that the number of customers is at least $l$, $l \geq 0$. The 'corresponding' $M/M/m$ queue is an ordinary $M/M/m$ queue with arrival rate $\lambda$ and service rate $\mu$.

For $l \leq m-1$,

$$\mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}}) = \left[\sum_{k=l}^{m-1} \frac{\rho^k}{k!} + \frac{\rho^m}{m!}\frac{m}{m-\rho}\right]^{-1} \left[\sum_{k=l}^{m-1} z^k \frac{\rho^k}{k!} + z^m \frac{\rho^m}{m!}\frac{m}{m-\rho z}\right]. \tag{2.1}$$

For $l \geq m-1$,

$$\mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}}) = z^l \frac{m-\rho}{m-\rho z}. \tag{2.2}$$

Lemma 2.1 expresses the distribution of $\mathbf{N}$ into the distribution of $\mathbf{N}_A$ and $\mathbf{N}^{(l)}_{M/M/m}$.

**Lemma 2.1**

$$\mathrm{E}(z^{\mathbf{N}}) = \gamma \left[\sum_{l=0}^{m-2} \frac{\mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}})\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}} + \frac{m}{m-\rho z} \sum_{l=m-1}^{\infty} z^l \Pr\{\mathbf{N}_A = l\}\right], \tag{2.3}$$

with

$$\gamma = \left[\sum_{l=0}^{m-2} \frac{\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}} + \frac{m}{m-\rho} \sum_{l=m-1}^{\infty} \Pr\{\mathbf{N}_A = l\}\right]^{-1}. \tag{2.4}$$

**Proof**

See appendix A.

□

**Remark 2.1** For $\Pr\{\mathbf{N}_A = 0\} = 1$, i.e., in a non-serving interval there are never any customers present, (2.3) and (2.4) reduce to

$$\mathrm{E}(z^{\mathbf{N}}) = \left[\sum_{l=0}^{m-1} \frac{\rho^l}{l!} + \frac{\rho^m}{m!}\frac{m}{m-\rho}\right]^{-1} \left[\sum_{l=0}^{m-1} z^l \frac{\rho^l}{l!} + z^m \frac{\rho^m}{m!}\frac{m}{m-\rho z}\right],$$

which is of course just the queue length distribution at an arbitrary epoch in the corresponding $M/M/m$ queue without service interruptions.

□

**Remark 2.2** For $m = 1$, (2.3) and (2.4) reduce to

$$\mathrm{E}(z^{\mathbf{N}}) = \frac{1-\rho}{1-\rho z}\mathrm{E}(z^{\mathbf{N}_A}),$$

which is the Fuhrmann-Cooper decomposition for an $M/M/1$ queue with service interruptions.
For $m = \infty$, (2.3) and (2.4) reduce to

$$\mathrm{E}(z^{\mathbf{N}}) = \left[\sum_{l=0}^{\infty}\Pr\{\mathbf{N}_A = l\}\frac{l!}{\rho^l}\sum_{k=l}^{\infty}\frac{\rho^k}{k!}\right]^{-1}\left[\sum_{l=0}^{\infty}\Pr\{\mathbf{N}_A = l\}\frac{l!}{\rho^l}\sum_{k=l}^{\infty}z^k\frac{\rho^k}{k!}\right].$$

Recognizing that $\dfrac{l!}{\rho^l}\displaystyle\sum_{k=l}^{\infty}z^k\frac{\rho^k}{k!} = z^l + \rho\int_{u=0}^{z}u^l\mathrm{e}^{(z-u)\rho}du,$

$$\mathrm{E}(z^{\mathbf{N}}) = \left[1+\rho\int_{u=0}^{1}u^l\mathrm{e}^{(1-u)\rho}du\right]^{-1}\left[\mathrm{E}(z^{\mathbf{N}_A}) + \rho\int_{u=0}^{z}\mathrm{E}(u^{\mathbf{N}_A})\mathrm{e}^{(z-u)\rho}du\right], \tag{2.5}$$

which will be useful later on.

□

Lemma 2.1 implies that to find the distribution of $\mathbf{N}$, it suffices to find the distribution of $\mathbf{N}_A$, as the distribution of $\mathbf{N}^{(l)}_{M/M/m}$ is known from (2.1). From a methodological point of view however it is more natural to analyze the queue length at either the beginning or the end of non-serving intervals than to study $\mathbf{N}_A$. Therefore we now relate the distribution of $\mathbf{N}_A$ to the queue length distribution at such embedded epochs. Denote by $\mathbf{N}_B^{(k)}$ and $\mathbf{N}_C^{(k)}$ the queue length at, respectively, the beginning and the end of the $k$-th non-serving interval. Denote by $\mathbf{N}_B$, $\mathbf{N}_C$ a pair of stochastic variables with as joint distribution the stationary joint distribution of $\mathbf{N}_B^{(k)}$, $\mathbf{N}_C^{(k)}$.

Lemma 2.2 relates the distribution of $\mathbf{N}_A$ to the distribution of $\mathbf{N}_B$ and $\mathbf{N}_C$.

**Lemma 2.2**

$$\Pr\{\mathbf{N}_A = l\} = \frac{\Pr\{\mathbf{N}_B \le l\} - \Pr\{\mathbf{N}_C \le l\}}{\mathrm{E}\mathbf{N}_C - \mathrm{E}\mathbf{N}_B}. \tag{2.6}$$

Written in terms of pgf's

$$\mathrm{E}(z^{\mathbf{N}_A}) = \frac{\mathrm{E}(z^{\mathbf{N}_B}) - \mathrm{E}(z^{\mathbf{N}_C})}{(1-z)(\mathrm{E}\mathbf{N}_C - \mathrm{E}\mathbf{N}_B)}. \tag{2.7}$$

**Proof**
Because of the PASTA property $\mathbf{N}_A$ has the same distribution as the number of customers seen by an arbitrary customer arriving in a non-serving interval. In the first $K$ non-serving intervals, the fraction of customers that see $l$ customers upon arrival is

$$\frac{\sum_{k=1}^{K} I_{\{\mathbf{N}_B^{(k)} \le l \le \mathbf{N}_C^{(k)}-1\}}}{\sum_{k=1}^{K} (\mathbf{N}_C^{(k)} - \mathbf{N}_B^{(k)})},$$

with $I_E$ denoting the indicator function of the event $E$. So, using the law of large numbers,

$$\begin{aligned} \Pr\{\mathbf{N}_A = l\} &= \lim_{K\to\infty} \frac{\sum_{k=1}^{K} I_{\{\mathbf{N}_B^{(k)} \le l \le \mathbf{N}_C^{(k)}-1\}}/K}{\sum_{k=1}^{K} (\mathbf{N}_C^{(k)} - \mathbf{N}_B^{(k)})/K} \\ &= \frac{\Pr\{\mathbf{N}_B \le l \le \mathbf{N}_C - 1\}}{\mathrm{E}\mathbf{N}_C - \mathrm{E}\mathbf{N}_B} = \frac{\Pr\{\mathbf{N}_B \le l\} - \Pr\{\mathbf{N}_C \le l\}}{\mathrm{E}\mathbf{N}_C - \mathrm{E}\mathbf{N}_B}, \end{aligned}$$

as $\Pr\{\mathbf{N}_B \le \mathbf{N}_C\} = 1$.

□

To illustrate Lemma 2.2 we now give two simple examples.

**Example 2.1**
Consider an exhaustive vacation system, where the servers together go on vacation when the system is empty. Then $\mathrm{E}(z^{\mathbf{N}_B}) = 1$. Let $\mathbf{V}$ be the length of an arbitrary vacation. Let $v(\omega) = \mathrm{E}(\mathrm{e}^{-\omega\mathbf{V}})$ for $\mathrm{Re}\,\omega \ge 0$. If the system is still empty when the servers return from vacation, then in the multiple vacation case the servers again go on vacation, i.e., $\mathrm{E}(z^{\mathbf{N}_C}) = \dfrac{v(\lambda(1-z)) - v(\lambda)}{1 - v(\lambda)}$, whereas in the single vacation case the servers just remain idling, awaiting a customer to arrive, i.e., $\mathrm{E}(z^{\mathbf{N}_C}) = v(\lambda(1-z)) - v(\lambda)(1-z)$. Summarizing, in the multiple vacation case

$$\mathrm{E}(z^{\mathbf{N}_A}) = \frac{1 - v(\lambda(1-z))}{(1-z)\lambda\mathrm{E}\mathbf{V}}, \tag{2.8}$$

whereas in the single vacation case

$$\mathrm{E}(z^{\mathbf{N}_A}) = \frac{1 - v(\lambda(1-z)) + v(\lambda)(1-z)}{(1-z)(\lambda\mathrm{E}\mathbf{V} + v(\lambda))}. \tag{2.9}$$

Browne & Kella [11] analyze an exhaustive $M/M/\infty$ system with multiple vacations. By a direct method they find $\mathrm{E}(z^{\mathbf{N}}) = H_{\mathbf{V}}(z)/H_{\mathbf{V}}(1)$ with

$$H_{\mathbf{V}}(z) = \frac{1 - v(\lambda(1-z))}{(1-z)\lambda\mathrm{E}\mathbf{V}} + \rho \int_{u=0}^{z} \frac{1 - v(\lambda(1-u))}{(1-u)\lambda\mathrm{E}\mathbf{V}} \mathrm{e}^{(z-u)\rho} du,$$

which agrees with (2.5), (2.8).

**Example 2.2**
Consider an exhaustive system with a stochastic **K**-policy, cf. Bisdikian [7], where service is interrupted when the system is empty, while service is resumed as soon as a stochastic number of **K** customers have accumulated again. Then

$$\mathrm{E}(z^{\mathbf{N}_A}) = \frac{1-\mathrm{E}(z^{\mathbf{K}})}{(1-z)\mathrm{E}\mathbf{K}}. \tag{2.10}$$

Browne & Kella [11] also study an exhaustive $M/M/\infty$ system with a deterministic $K$-policy. By a direct method they find $\mathrm{E}(z^{\mathbf{N}}) = H_K(z)/H_K(1)$ with

$$H_K(z) = \frac{1-z^K}{1-z} + \rho \sum_{k=0}^{K-1} \int_{u=0}^{z} u^k \mathrm{e}^{(z-u)\rho} du,$$

which agrees with (2.5), (2.10).

We now show how the queue length decomposition translates into a decomposition of the waiting time. In addition to (i) and (ii) we make the following assumptions.
(iii). Customers enter service in order of arrival.
(iv). The waiting time of customers is independent of arrivals after their own arrival.
Denote by **W** and **R** respectively the waiting and the sojourn time of an arbitrary customer. Denote by **L** the number of waiting customers at an arbitrary epoch. The familiar relationship $\mathrm{E}(z^{\mathbf{N}}) = \mathrm{E}(\mathrm{e}^{-\lambda(1-z)\mathbf{R}})$ does *not* hold here, as customers do not necessarily leave in order of arrival. However, what *does* hold under the assumptions (iii) and (iv) is the relationship $\mathrm{E}(z^{\mathbf{L}}) = \mathrm{E}(\mathrm{e}^{-\lambda(1-z)\mathbf{W}})$. What thus remains to be done, is to relate the distribution of **L** to the distribution of **N**.

**Lemma 2.3**

$$\mathrm{E}(z^{\mathbf{L}}) = \mathrm{E}(z^{\mathbf{N}}) + (1-p_A)\left([z^{-m}-1]\mathrm{E}(z^{\mathbf{N}_E}) + \sum_{k=0}^{m-1}[1-z^{k-m}]\Pr\{\mathbf{N}_E = k\}\right), \tag{2.11}$$

with

$$\mathrm{E}(z^{\mathbf{N}_E}) = \frac{\mathrm{E}(z^{\mathbf{N}}) - p_A\mathrm{E}(z^{\mathbf{N}_A})}{1-p_A}, \qquad p_A = \left[\sum_{l=0}^{\infty} \frac{\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}\right]^{-1}. \tag{2.12}$$

**Proof**
Denote by $\mathbf{N}_E$ the number of customers present at an arbitrary epoch in a serving interval. Denote by $p_A$ the fraction of time occupied by non-serving intervals. Then

$$\mathrm{E}(z^{\mathbf{N}}) = \mathrm{E}(z^{\mathbf{N}_A})p_A + \mathrm{E}(z^{\mathbf{N}_E})(1-p_A), \tag{2.13}$$

$$\mathrm{E}(z^{\mathbf{L}}) = \mathrm{E}(z^{\mathbf{N}_A})p_A + \mathrm{E}(z^{[\mathbf{N}_E-m]^+})(1-p_A), \tag{2.14}$$

with $[x]^+ = \max(0,x)$.
Comparing (2.13), (2.14), using that

$$\mathrm{E}(z^{[\mathbf{N}_E-m]^+}) = z^{-m}\mathrm{E}(z^{\mathbf{N}_E}) + \sum_{k=0}^{m-1}[1-z^{k-m}]\Pr\{\mathbf{N}_E = k\} \tag{2.15}$$

yields (2.11).
Because of the PASTA property $p_A$ equals the probability that an arbitrary customer arrives in a non-serving interval. In the proof of Lemma 2.1 we introduced the notion of a 1-busy period. We showed that in a 1-busy period exactly 1 customer is served that arrived in a non-serving interval. Denote by $\mathbf{M}$ the number of customers served in a 1-busy period. Then

$$p_A = \frac{1}{\mathrm{EM}}. \tag{2.16}$$

From the proof of Lemma 2.1

$$\mathrm{EM} = \sum_{l=0}^{\infty} \frac{\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}. \tag{2.17}$$

Substituting (2.17) into (2.16) completes the proof.

□

## 3 The joint queue length distribution I

We now return to the polling system with multiple coupled servers. We first present a detailed model description. The model under consideration consists of $n$ queues $Q_1, \ldots, Q_n$, each of infinite capacity, attended to by $m$ coupled servers. Customers arrive at the queues according to independent Poisson processes. Customers arriving at $Q_i$ will also be referred to as type-$i$ customers, $i = 1, \ldots, n$. Denote by $\lambda_i$ the arrival rate at $Q_i$, $i = 1, \ldots, n$. The total arrival rate is $\lambda := \sum_{i=1}^{n} \lambda_i$. Type-$i$ customers require service times $\mathbf{B}_i$, having distribution function $B_i(\cdot)$ with LST $\beta_i(\cdot)$ and first moment $\beta_i$, $i = 1, \ldots, n$. Define the traffic intensity at $Q_i$ as $\rho_i := \lambda_i\beta_i$, $i = 1, \ldots, n$. The total traffic intensity is $\rho := \sum_{i=1}^{n} \rho_i$. The server pool visits the queues in strictly cyclic order $Q_1, \ldots, Q_n$. Moving from $Q_i$ to $Q_{i+1}$, the server pool experiences a non-zero switch-over time $\mathbf{S}_i$, having distribution function $S_i(\cdot)$ with LST $\sigma_i(\cdot)$ and first moment $s_i$, $i = 1, \ldots, n$. Here (as well as in the sequel) $n+1$ is to be understood as 1. Successive service times as well as successive switch-over times are assumed to be independent. Also the arrival process, the service process, and the switch-over process are assumed to be mutually independent. As soon as the servers arrive at $Q_i$, they start serving type-$i$ customers, as prescribed by the service discipline. For now we do not specify the service discipline any further. In fact, what we are mainly interested in, is exploring the class of service disciplines that allow an exact analysis. As soon as the servers have finished serving type-$i$ customers, as prescribed by the service discipline, they move to $Q_{i+1}$.
In the present section we relate the pgf of the joint queue length distribution at the beginning of a visit to $Q_i$ to the pgf of the joint queue length distribution at the end of a visit to $Q_{i-1}$. Next we also relate the pgf of the joint queue length distribution at the *end* of a visit to $Q_i$ to the pgf of the joint queue length distribution at the *beginning* of a visit to $Q_i$. Thus we obtain $2n$ equations involving $2n$ pgf's. In the next section we identify some cases in which these pgf's can actually be solved from these equations.

We first introduce some notation. Denote by $\mathbf{X}_{ih}$ and $\mathbf{Y}_{ih}$ stochastic variables with distribution the stationary queue length distribution at $Q_h$ at, respectively, the beginning and the end of a visit to $Q_i$, $h = 1, \ldots, n$, $i = 1, \ldots, n$.
Define

$$F_i(z) = \mathrm{E}(z_1^{\mathbf{X}_{i1}} \ldots z_n^{\mathbf{X}_{in}})$$

$$G_i(z) = \mathrm{E}(z_1^{\mathbf{Y}_{i1}} \ldots z_n^{\mathbf{Y}_{in}})$$

for $z = (z_1, \ldots, z_n)$, $|z_h| \leq 1$, $h = 1, \ldots, n$, $i = 1, \ldots, n$.
We first relate $F_i(\cdot)$ to $G_{i-1}(\cdot)$, and subsequently $G_i(\cdot)$ to $F_i(\cdot)$. Thus we obtain an expression for $G_i(\cdot)$ in terms of $G_{i-1}(\cdot)$, which recursively yields a functional equation for $G_i(\cdot)$.
Define

$$d_i(z) = \sigma_i(\sum_{h=1}^{n} \lambda_h(1 - z_h)) \tag{3.1}$$

for $z = (z_1, \ldots, z_n)$, $|z_h| \leq 1$, $h = 1, \ldots, n$, $i = 1, \ldots, n$.
Then

$$F_i(z) = G_{i-1}(z)d_{i-1}(z), \tag{3.2}$$

where $d_0(\cdot)$, $G_0(\cdot)$ are to be understood as $d_n(\cdot)$, $G_n(\cdot)$ respectively.

**Remark 3.1** In accordance with the model description, we assume here that the switch-over times are non-zero and that the servers keep switching when the system is empty. In case the switch-over times are zero, or in case the servers stop switching when the system is empty, (3.2) should be modified into

$$F_i(z) = G_{i-1}(z)d_{i-1}(z) + G_{i-1}(0)d_{i-1}(z)[e_i(z) - 1]$$

with

$$e_i(z) = \sum_{h \neq i} \frac{\lambda_h}{\lambda} z_h + \frac{\lambda_i}{\lambda} \theta_i(z).$$

Here $\theta_i(z)$ depends on what happens when an arriving type-$i$ customer sees the servers idling at $Q_i$.

□

We now relate $G_i(\cdot)$ to $F_i(\cdot)$.

$$G_i(z) = \sum_{l_1=0}^{\infty} \ldots \sum_{l_n=0}^{\infty} \mathrm{E}(z_1^{\mathbf{Y}_{i1}} \ldots z_n^{\mathbf{Y}_{in}} \mid (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in}) = (l_1, \ldots, l_n)) \tag{3.3}$$
$$\Pr\{(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in}) = (l_1, \ldots, l_n)\}.$$

Evidently, it is the service discipline at $Q_i$ that decides whether or not the right-hand side of (3.3) can be expressed into $F_i(\cdot)$. Fuhrmann [13] and Resing [26] consider a class of service disciplines (in single-server systems) that satisfy the following assumption.

**Assumption 3.1**
If there are $l_i$ customers present at $Q_i$ at the start of a visit, then during the course of the visit each of these $l_i$ customers will effectively be replaced in an i.i.d. manner by a random population consisting of $\mathbf{K}_{i1}$ type-1 customers, ..., $\mathbf{K}_{in}$ type-$n$ customers having $n$-dimensional pgf $\eta_i(z) = \mathrm{E}(z_1^{\mathbf{K}_{i1}} \dots z_n^{\mathbf{K}_{in}})$.
Formally,

$$\mathrm{E}(z_1^{\mathbf{Y}_{i1}} \dots z_n^{\mathbf{Y}_{in}} \mid (\mathbf{X}_{i1}, \dots, \mathbf{X}_{in}) = (l_1, \dots, l_n)) = z_1^{l_1} \dots z_{i-1}^{l_{i-1}} (\eta_i(z))^{l_i} z_{i+1}^{l_{i+1}} \dots z_n^{l_n}. \quad (3.4)$$

Substituting (3.4) into (3.3),

$$G_i(z) = F_i(z_1, \dots, z_{i-1}, \eta_i(z), z_{i+1}, \dots, z_n). \quad (3.5)$$

Using the theory of multi-type branching processes, both Fuhrmann and Resing show that the class of service disciplines that satisfy Assumption 3.1, like exhaustive and gated, allows a relatively simple exact analysis, basically due to the relatively simple form of (3.5). The results suggest that service disciplines that violate Assumption 3.1 defy an exact analysis, except for some special cases like two-queue cases and completely symmetrical cases.

In multiple-server systems there are no non-trivial service disciplines that satisfy Assumption 3.1. However, some service disciplines *do* satisfy the following somewhat milder assumption than Assumption 3.1.

**Assumption 3.2**
If there are $l_i$ customers present at $Q_i$ at the start of a visit, then during the course of the visit one of these $l_i$ customers will effectively be replaced by a random population having pgf $\eta_i^{(1)}(z)$, while each of the other customers will effectively be replaced in an i.i.d. manner by a random population having pgf $\eta_i(z)$.
Formally,

$$\mathrm{E}(z_1^{\mathbf{Y}_{i1}} \dots z_n^{\mathbf{Y}_{in}} \mid (\mathbf{X}_{i1}, \dots, \mathbf{X}_{in}) = (l_1, \dots, l_n)) = z_1^{l_1} \dots z_{i-1}^{l_{i-1}} \eta_i^{(l_i)}(z) z_{i+1}^{l_{i+1}} \dots z_n^{l_n}, \quad (3.6)$$

with

$$\eta_i^{(l_i)}(z) = 1, \qquad l_i = 0, \quad (3.7)$$

$$\eta_i^{(l_i)}(z) = \eta_i^{(1)}(z)(\eta_i(z))^{l_i - 1}, \qquad l_i > 0. \quad (3.8)$$

Below we will describe some multiple-server systems with service disciplines that satisfy Assumption 3.2. We will then also briefly indicate some circumstances that may occur in single-server systems under the influence of which in principle simple service disciplines violate Assumption 3.1 but still satisfy Assumption 3.2.
Substituting (3.6), (3.7), (3.8) into (3.3),

$$\begin{aligned} G_i(z) &= F_i(z_1, \dots, z_{i-1}, \eta_i(z), z_{i+1}, \dots, z_n) \frac{\eta_i^{(1)}(z)}{\eta_i(z)} \\ &+ F_i(z_1, \dots, z_{i-1}, 0, z_{i+1}, \dots, z_n)[1 - \frac{\eta_i^{(1)}(z)}{\eta_i(z)}]. \end{aligned} \quad (3.9)$$

Define

$$a_i(z) = (z_1, \ldots, z_{i-1}, \eta_i(z), z_{i+1}, \ldots, z_n); \tag{3.10}$$

$$b_i(z) = (z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_n); \tag{3.11}$$

$$c_i(z) = \frac{\eta_i^{(1)}(z)}{\eta_i(z)} \tag{3.12}$$

for $z = (z_1, \ldots, z_n)$, $|z_h| \leq 1$, $h = 1, \ldots, n$, $i = 1, \ldots, n$.
Then (3.9) may be written as

$$G_i(z) = F_i(a_i(z))c_i(z) + F_i(b_i(z))[1 - c_i(z)]. \tag{3.13}$$

In view of the results of Fuhrmann and Resing, one can in general not expect that the class of service disciplines that satisfy Assumption 3.2 but not Assumption 3.1, i.e., with $c_i(z) \neq 1$, allows an exact analysis, except possibly for some special cases. In the next section we will identify some of those cases.

We now describe some multiple-server systems with service disciplines that satisfy Assumption 3.2. Assumption 3.2 says that during the course of a visit to $Q_i$ one of the customers initially present gets replaced by a different population than all the others. This suggests that either only one of the customers initially present at $Q_i$ actually gets served or that all of them get served but that one of them keeps the servers busy for a different time than all the others. Having this in mind, we consider a class of service disciplines that are parametrized by a vector $(L_1, \ldots, L_n)$ of integers and a vector $(p_1, \ldots, p_n)$ of probabilities with the following interpretation. If there are $l_i$ customers present at the start of a visit to $Q_i$, then $k_i = \min(l_i, L_i)$ of them are served. Customers arriving at $Q_i$ during the course of a visit, are served with probability $p_i$. The case $L_i = 1$ contains both the semi-exhaustive service discipline ($p_i = 1$) and the 1-limited service discipline ($p_i = 0$). The case $L_i = \infty$ includes both the exhaustive service discipline ($p_i = 1$) and the gated service discipline ($p_i = 0$).
Denote $\kappa_i = \lambda_i p_i$.
Let $\mathbf{T}_i^{(k)}$ be the length of a busy period starting with $k$ customers present in an ordinary $M/G/m$ queue with arrival rate $\kappa_i$ and service time distribution $B_i(\cdot)$. Let $\tau_i^{(k)}(\omega) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{T}_i^{(k)}})$ for $\mathrm{Re}\,\omega \geq 0$.
Define

$$\alpha_i(z) = \sum_{h \neq i} \lambda_h(1 - z_h) + \lambda_i(1 - p_i)(1 - z_i)$$

for $z = (z_1, \ldots, z_n)$, $|z_h| \leq 1$, $h = 1, \ldots, n$, $i = 1, \ldots, n$.
For the above-defined class of service disciplines

$$\eta_i^{(l_i)}(z) = z_i^{l_i - k_i} \tau_i^{(k_i)}(\alpha_i(z)), \tag{3.14}$$

with $k_i = \min(l_i, L_i)$ and the interpretation of $\eta_i^{(l_i)}(z)$ as in (3.8). For $L_i = 1$ (3.14) satisfies (3.7) and (3.8) in Assumption 3.2 with $\eta_i(z) = z_i$, $\eta_i^{(1)}(z) = \tau_i^{(1)}(\alpha_i(z))$. If $\tau_i^{(k_i)}(\cdot)$ is of the form

$$\tau_i^{(k_i)}(\omega) = 1, \qquad k_i = 0, \tag{3.15}$$

$$\tau_i^{(k_i)}(\omega) = \tau_i^{(1)}(\omega)(\tau_i(\omega))^{k_i-1}, \qquad k_i > 0, \tag{3.16}$$

for some LST $\tau_i(\cdot)$, then (3.14) satisfies (3.7) and (3.8) also for $L_i = \infty$ with $\eta_i(z) = \tau_i(\alpha_i(z))$, $\eta_i^{(1)}(z) = \tau_i^{(1)}(\alpha_i(z))$.
We now give two examples where $\tau_i^{(k_i)}(\cdot)$ is of the form (3.15), (3.16).

**Example 3.1**
Assume that the service times at $Q_i$ are exponentially distributed with parameter $\mu_i = 1/\beta_i$. Let $\mathbf{U}_i^{(k,l)}$ be the time needed in an $M/M/m$ queue with arrival rate $\kappa_i$ and service rate $\mu_i$ to reduce the queue length from $k$ to $l$, $k \geq l$. Let $\phi_i^{(k,l)}(\omega) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{U}_i^{(k,l)}})$ for $\mathrm{Re}\,\omega \geq 0$, $k \geq l$. So $\tau_i^{(k_i)}(\omega) = \phi_i^{(k_i,0)}(\omega)$.
Let $\mathbf{U}_i$ be the length of a busy period in an $M/M/1$ queue with arrival rate $\kappa_i$ and service rate $m\mu_i$. Let $\phi_i(\omega) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{U}_i})$ for $\mathrm{Re}\,\omega \geq 0$.
Because of the properties of the exponential distribution

$$\phi_i^{(k,0)}(\omega) = \prod_{l=1}^{k} \phi_i^{(l,l-1)}(\omega),$$

$$\phi_i^{(l,l-1)}(\omega) = \phi_i(\omega), \qquad l \geq m.$$

The transforms $\phi_i(\cdot)$, $\phi_i^{(1,0)}(\cdot), \ldots, \phi_i^{(m-1,m-2)}(\cdot)$ may be determined by solving a set of linear equations, cf. Medhi [21] pp. 138-140. In particular,

$$\phi_i(\omega) = \frac{\kappa_i + m\mu_i - \sqrt{(\kappa_i + \omega + m\mu_i)^2 - 4m\kappa_i\mu_i}}{2\kappa_i},$$

$$\phi_i^{(m-1,m-2)}(\omega) = \frac{(m-1)\phi_i(\omega)}{m - \phi_i(\omega)}.$$

For the gated service discipline, i.e., $p_i = 0$,

$$\phi_i(\omega) = \frac{m\mu_i}{m\mu_i + \omega},$$

$$\phi_i^{(l,l-1)}(\omega) = \frac{l\mu_i}{l\mu_i + \omega}, \qquad l \leq m.$$

Summarizing, for $m = 1$, $\tau_i^{(k_i)}(\cdot)$ is of the form (3.15), (3.16) with $\tau_i(\omega) = \tau_i^{(1)}(\omega) = \phi_i(\omega)$. Also for $m = 2$, $\tau_i^{(k_i)}(\cdot)$ is of the form (3.15), (3.16) with $\tau_i(\omega) = \phi_i(\omega)$, $\tau_i^{(1)}(\omega) = \phi_i^{(1,0)}(\omega)$, $\phi_i^{(1,0)}(\omega) = \dfrac{\phi_i(\omega)}{2 - \phi_i(\omega)}$. For $m > 2$, $\tau_i^{(k_i)}(\cdot)$ is no longer of the form (3.15), (3.16).

**Example 3.2**
Assume that the service times at $Q_i$ are deterministic.
Let $\mathbf{V}_i$ be the length of a busy period in an ordinary $M/D/m$ queue with arrival rate $\kappa_i$ and service time $\beta_i$. Let $\psi_i(\omega) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{V}_i})$ for $\mathrm{Re}\,\omega \geq 0$.
For $m = \infty$, $\tau_i^{(k_i)}(\cdot)$ is of the form (3.15), (3.16) with $\tau_i(\omega) = 1$, $\tau_i^{(1)}(\omega) = \psi_i(\omega)$,
$\psi_i(\omega) = \dfrac{(\omega + \lambda_i)\mathrm{e}^{-\beta_i(\omega+\lambda_i)}}{\omega + \lambda_i \mathrm{e}^{-\beta_i(\omega+\lambda_i)}}$, cf. Stadje [27].

**Remark 3.2** There are also some circumstances that may occur in single-server systems under the influence of which in principle simple service disciplines violate Assumption 3.1 but still satisfy Assumption 3.2. An obvious example arises when one of the customers served during a visit, e.g. the first one or the last one, requires an exceptional service time. Another example is provided by a set-up time or a shut-down time that is only incurred when at least one customer is served during a visit.

□

## 4 THE JOINT QUEUE LENGTH DISTRIBUTION II

In the previous section we obtained under Assumption 3.2 a set of $2n$ equations (3.2), (3.13) involving the $2n$ pgf's $F_i(z)$, $G_i(z)$, $i = 1, \ldots, n$. In the present section we identify some cases in which these pgf's can actually be solved from these equations. Obviously it suffices to find either $F_i(z)$ or $G_i(z)$ for an arbitrary $i$, as the remaining $F_i(z)$, $G_i(z)$, $i = 1, \ldots, n$, can then easily be found from (3.2), (3.13).
Substituting (3.2) into (3.13),

$$G_i(z) = G_{i-1}(a_i(z))d_{i-1}(a_i(z))c_i(z) + G_{i-1}(b_i(z))d_{i-1}(b_i(z))[1 - c_i(z)], \tag{4.1}$$

where $d_0(\cdot)$, $G_0(\cdot)$ are to be understood as $d_n(\cdot)$, $G_n(\cdot)$ respectively.
Applying (4.1) $n$ times we obtain a functional equation for $G_i(\cdot)$.
For $n = 1$ we find, using the definitions (3.1), (3.10), (3.11), (3.12),

$$G(z) = G(\eta(z))\sigma(\lambda(1 - \eta(z)))\frac{\eta^{(1)}(z)}{\eta(z)} + G(0)\sigma(\lambda)[1 - \frac{\eta^{(1)}(z)}{\eta(z)}]. \tag{4.2}$$

Here (as well as in the sequel) the redundant indices are omitted.
For $n = 2$ we find

$$\begin{aligned} G_i(z) &= G_i(a_{i-1}(a_i(z)))d_{i-1}(a_{i-1}(a_i(z)))c_{i-1}(a_i(z))d_i(a_i(z))c_i(z) \\ &+ G_i(b_{i-1}(a_i(z)))d_{i-1}(b_{i-1}(a_i(z)))[1 - c_{i-1}(a_i(z))]d_i(a_i(z))c_i(z) \\ &+ G_i(a_{i-1}(b_i(z)))d_{i-1}(a_{i-1}(b_i(z)))c_{i-1}(b_i(z))d_i(b_i(z))[1 - c_i(z)] \\ &+ G_i(b_{i-1}(b_i(z)))d_{i-1}(b_{i-1}(b_i(z)))[1 - c_{i-1}(b_i(z))]d_i(b_i(z))[1 - c_i(z)]. \end{aligned}$$

Using the definitions (3.1), (3.10), (3.11), (3.12),

$$G_1(z_1,z_2) = \tag{4.3}$$

$$G_1(\eta_1(z),\eta_2(\eta_1(z),z_2))\sigma_2\{\gamma(\eta_1(z),\eta_2(\eta_1(z),z_2))\}\sigma_1\{\gamma(\eta_1(z),z_2)\}\frac{\eta_2^{(1)}(\eta_1(z),z_2)}{\eta_2(\eta_1(z),z_2)}\frac{\eta_1^{(1)}(z)}{\eta_1(z)}+$$

$$G_1(\eta_1(z),0)\sigma_2\{\gamma(\eta_1(z),0)\}\sigma_1\{\gamma(\eta_1(z),z_2)\}[1-\frac{\eta_2^{(1)}(\eta_1(z),z_2)}{\eta_2(\eta_1(z),z_2)}]\frac{\eta_1^{(1)}(z)}{\eta_1(z)}+$$

$$G_1(0,\eta_2(0,z_2))\sigma_2\{\gamma(0,\eta_2(0,z_2))\}\sigma_1\{\gamma(0,z_2)\}\frac{\eta_2^{(1)}(0,z_2)}{\eta_2(0,z_2)}[1-\frac{\eta_1^{(1)}(z)}{\eta_1(z)}]+$$

$$G_1(0,0)\sigma_2\{\gamma(0,0)\}\sigma_1\{\gamma(0,z_2)\}[1-\frac{\eta_2^{(1)}(0,z_2)}{\eta_2(0,z_2)}][1-\frac{\eta_1^{(1)}(z)}{\eta_1(z)}],$$

(similarly with the indices interchanged) with $\gamma(z_1,z_2)=\lambda_1(1-z_1)+\lambda_2(1-z_2)$. Remember that $\eta_i(z)$ is an $n$-dimensional pgf so that $\mid \eta_i(z) \mid \leq 1$ for $z=(z_1,\ldots,z_n)$, $\mid z_h \mid \leq 1$, $h=1,\ldots,n$, $i=1,\ldots,n$.
In general we obtain a functional equation for $G_i(\cdot)$ containing $2^n$ arguments in the right-hand side. So, in accordance with the results of Fuhrmann and Resing, in general the functional equation can not be solved. In fact solving the functional equation only stands a chance in cases where 'enough' of the $2^n$ arguments in the right-hand side reduce either to $z$ or to a constant. We will now indicate some of those cases.

*Case I. $n=1$ queue, $\eta(z)=z$.*
This covers the case $L=1$ described in the previous section, i.e., only one of the customers present at the start of a visit is served, while customers arriving during the course of a visit are served with probability $p$.
Rewriting (4.2),

$$G(z)[z-\sigma(\lambda(1-z))\eta^{(1)}(z)] = G(0)\sigma(\lambda)[z-\eta^{(1)}(z)]. \tag{4.4}$$

Letting $z \to 1$ in (4.4),

$$G(0) = \frac{1}{\sigma(\lambda)}\frac{1-(\eta^{(1)})'(1)-\lambda s}{1-(\eta^{(1)})'(1)},$$

with $(\eta^{(1)})'(1) = \dfrac{d\eta^{(1)}(z)}{dz}_{|z=1}$. Apparently the stability condition is $\lambda s+(\eta^{(1)})'(1)<1$. Note that $\lambda s+(\eta^{(1)})'(1)$ is the mean increase of the queue length between the start of two successive visits when the system is not empty, which should indeed be less than 1 to ensure stability.

*Case II. $n=1$ queue, $\eta(z) \neq z$.*
This covers the case $L=\infty$ described in the previous section, i.e., all the customers present at the start of a visit are served, while customers arriving during the course of a visit are served with probability $p$, moreover assuming that there are either two servers and exponential service times, cf. Example 3.1, or an infinite number of servers and deterministic service times, cf. Example 3.2.

Writing $e(z) = \eta(z)$, $f(z) = \sigma(\lambda(1-\eta(z)))\eta^{(1)}(z)/\eta(z)$, $g(z) = \sigma(\lambda)[1-\eta^{(1)}(z)/\eta(z)]$ in (4.2),

$$G(z) = G(e(z))f(z) + G(0)g(z). \tag{4.5}$$

Define

$$e^{(0)}(z) = z; \qquad e^{(k)}(z) = e(e^{(k-1)}(z)); \qquad k \geq 1,$$

for $|z| \leq 1$.
Iterating (4.5) $K$ times,

$$G(z) = G(e^{(K+1)}(z)) \prod_{k=0}^{K} f(e^{(k)}(z)) + G(0) \sum_{k=0}^{K} g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z)). \tag{4.6}$$

The next Lemma establishes the convergence of (4.6) for $K \to \infty$ under the condition $\eta'(1) < 1$.

**Lemma 4.1**
If $\eta'(1) < 1$ then
*i.* $\lim_{K\to\infty} e^{(K+1)}(z) = 1$ for all $z$ with $|z| \leq 1$;
*ii.* $\prod_{k=0}^{\infty} f(e^{(k)}(z))$ converges for all $z$ with $|z| \leq 1$;
*iii.* $\sum_{k=0}^{\infty} g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z))$ converges for all $z$ with $|z| \leq 1$.

**Proof**
See appendix B.

□

Apparently the stability condition is $\eta'(1) < 1$. Note that $\eta'(1)$ is the mean number of customers by which each of the customers present at the start of a visit, except one, gets replaced in the course of the visit, which should indeed be less than 1 to ensure stability. In Example 3.1, $\eta'(1) = \dfrac{(1-p)\rho}{2-p\rho} < 1$ iff $\rho < 2$, irrespective of $p$. In Example 3.2, $\eta'(1) = 0$, also irrespective of $p$.
If $\eta'(1) < 1$ then, letting $K \to \infty$ in (4.6),

$$G(z) = \prod_{k=0}^{\infty} f(e^{(k)}(z)) + G(0) \sum_{k=0}^{\infty} g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z)). \tag{4.7}$$

Putting $z = 0$ in (4.7),

$$G(0) = \frac{\prod_{k=0}^{\infty} f(e^{(k)}(0))}{1 - \sum_{k=0}^{\infty} g(e^{(k)}(0)) \prod_{l=0}^{k-1} f(e^{(l)}(0))}.$$

*Case III.* $n = 2$ *queues,* $\eta_i(z) = z_i$, $i = 1, 2$.
This covers the case $L_i = 1$ described in the previous section, i.e., only one of the customers present at the start of a visit to $Q_i$ is served, while customers arriving at $Q_i$ during the course of a visit are served with probability $p_i$, $i = 1, 2$.
Equation (4.3) reduces to

$$G_1(z_1, z_2)[z_1 z_2 - \sigma_1(\gamma(z))\sigma_2(\gamma(z))\eta_1^{(1)}(z)\eta_2^{(1)}(z)] =$$

$$G_1(z_1, 0)\sigma_1(\gamma(z))\sigma_2(\gamma(z_1, 0))\eta_1^{(1)}(z)[z_2 - \eta_2^{(1)}(z)] +$$

$$G_1(0, z_2)\sigma_1(\gamma(0, z_2))\sigma_2(\gamma(0, z_2))[z_1 - \eta_1^{(1)}(z)]\eta_2^{(1)}(0, z_2) +$$

$$G_1(0, 0)\sigma_1(\gamma(0, z_2))\sigma_2(\gamma(0))[z_1 - \eta_1^{(1)}(z)][z_2 - \eta_2^{(1)}(0, z_2)].$$

For $p_i = 0$, i.e., $\eta_i^{(1)}(z) = \beta_i(\gamma(z))$, $i = 1, 2$, the problem of solving the above functional equation may be formulated as a boundary value problem, cf. Boxma & Groenendijk [9].

*Case IV.* $n = 2$ *queues,* $\eta_i(z)$, $\eta_i^{(1)}(z)$ *do not depend on* $z_i$, $i = 1, 2$.
This occurs in Example 3.1 for $p_i = 1$, $i = 1, 2$, i.e., two servers, exponential service times, exhaustive service.
If $\eta_i(z)$, $\eta_i^{(1)}(z)$ do not depend on $z_i$, $i = 1, 2$, then the complete right-hand side of (4.3) does not depend on $z_i$. In other words, $G_i(z)$ does not depend on $z_i$, reflecting that $Q_i$ is empty at the completion of a visit to $Q_i$ when $p_i = 1$. So equation (4.3) may be replaced by

$$H_1(z_2) = H_1(e_1(z_2))f_1(z_2) + H_1(\eta_2(0))g_1(z_2) + H_1(0)h_1(z_2), \tag{4.8}$$

with

$$\begin{aligned}
e_1(z_2) &= \eta_2(\eta_1(z), z_2); \\
f_1(z_2) &= \sigma_2\{\gamma(\eta_1(z), \eta_2(\eta_1(z), z_2))\}\sigma_1\{\gamma(\eta_1(z), z_2)\}\frac{\eta_2^{(1)}(\eta_1(z), z_2)}{\eta_2(\eta_1(z), z_2)}\frac{\eta_1^{(1)}(z)}{\eta_1(z)}; \\
g_1(z_2) &= \sigma_2\{\gamma(0, \eta_2(0, z_2))\}\sigma_1\{\gamma(0, z_2)\}\frac{\eta_2^{(1)}(0, z_2)}{\eta_2(0, z_2)}[1 - \frac{\eta_1^{(1)}(z)}{\eta_1(z)}]; \\
h_1(z_2) &= \sigma_2\{\gamma(\eta_1(z), 0)\}\sigma_1\{\gamma(\eta_1(z), z_2)\}[1 - \frac{\eta_2^{(1)}(\eta_1(z), z_2)}{\eta_2(\eta_1(z), z_2)}]\frac{\eta_1^{(1)}(z)}{\eta_1(z)} \\
&+ \sigma_2\{\gamma(0, 0)\}\sigma_1\{\gamma(0, z_2)\}[1 - \frac{\eta_2^{(1)}(0, z_2)}{\eta_2(0, z_2)}][1 - \frac{\eta_1^{(1)}(z)}{\eta_1(z)}]; \\
H_1(z_2) &= G_1(z_1, z_2).
\end{aligned}$$

Define

$$e_1^{(0)}(y) = y; \qquad e_1^{(k)}(y) = e_1(e_1^{(k-1)}(y)); \qquad k \geq 1,$$

for $|y| \leq 1$.

Iterating (4.8) $K$ times, writing $z_2 = y$,

$$\begin{aligned} H_1(y) &= H_1(e_1^{(K+1)}(y)) \prod_{k=0}^{K} f_1(e_1^{(k)}(y)) \\ &+ H_1(\eta_2(0)) \sum_{k=0}^{K} g_1(e_1^{(k)}(y)) \prod_{l=0}^{k-1} f_1(e_1^{(l)}(y)) \\ &+ H_1(0) \sum_{k=0}^{K} h_1(e_1^{(k)}(y)) \prod_{l=0}^{k-1} f_1(e_1^{(l)}(y)). \end{aligned} \tag{4.9}$$

The next Lemma establishes the convergence of (4.9) for $K \to \infty$ under the condition $e_1'(1) < 1$.

**Lemma 4.2**
If $e_1'(1) < 1$ then

*i.* $\lim_{K\to\infty} e_1^{(K+1)}(y) = 1$ for all $y$ with $|y| \leq 1$;

*ii.* $\prod_{k=0}^{\infty} f_1(e_1^{(k)}(y))$ converges for all $y$ with $|y| \leq 1$;

*iii.* $\sum_{k=0}^{\infty} g_1(e_1^{(k)}(y)) \prod_{l=0}^{k-1} f_1(e_1^{(l)}(y))$ converges for all $y$ with $|y| \leq 1$;

*iv.* $\sum_{k=0}^{\infty} h_1(e_1^{(k)}(y)) \prod_{l=0}^{k-1} f_1(e_1^{(l)}(y))$ converges for all $y$ with $|y| \leq 1$.

**Proof**
Similar to the proof of Lemma 4.1.

□

Apparently the stability condition is $e_1'(1) < 1$. Note that $e_1'(1)$ is the mean number of type-1 customers by which each of the type-1 customers present at the start of a cycle, except one, gets replaced during the course of the cycle. In Example 3.1 when $p_i = 1$, $e_1'(1) = \eta_1'(1)\eta_2'(1) = \dfrac{\rho_1}{2-\rho_1}\dfrac{\rho_2}{2-\rho_2} = \dfrac{\rho_1\rho_2}{4-2\rho+\rho_1\rho_2} < 1$ iff $\rho < 2$.
If $e_1'(1) < 1$ then, letting $K \to \infty$ in (4.9),

$$\begin{aligned} H_1(y) &= \prod_{k=0}^{\infty} f_1(e_1^{(k)}(y)) \\ &+ H_1(\eta_2(0)) \sum_{k=0}^{\infty} g_1(e_1^{(k)}(y)) \prod_{l=0}^{k-1} f_1(e_1^{(l)}(y)) \\ &+ H_1(0) \sum_{k=0}^{\infty} h_1(e_1^{(k)}(y)) \prod_{l=0}^{k-1} f_1(e_1^{(l)}(y)). \end{aligned} \tag{4.10}$$

Putting $y = 0$ and $y = \eta_2(0)$ in (4.10) we obtain a pair of linear equations for the unknown constants $H_1(0)$ and $H_1(\eta_2(0))$.

*Case V.* $n = 2$ *queues,* $\eta_i(z) = 1$, $i = 1, 2$.
This occurs in Example 3.2, i.e., an infinite number of servers, deterministic service times, all the customers present at the start of a visit are served, while customers arriving during the course of a visit are served with probability $p_i$, $i = 1, 2$.
Equation (4.3) reduces to

$$\begin{aligned} G_1(z_1, z_2) &= \sigma_1(\lambda_2(1 - z_2))\eta_2^{(1)}(1, z_2)\eta_1^{(1)}(z) \\ &+ G_1(1, 0)\sigma_2(\lambda_2)\sigma_1(\lambda_2(1 - z_2))[1 - \eta_2^{(1)}(1, z_2)]\eta_1^{(1)}(z) \\ &+ G_1(0, 1)\sigma_2(\lambda_1)\sigma_1(\lambda_1 + \lambda_2(1 - z_2))\eta_2^{(1)}(0, z_2)[1 - \eta_1^{(1)}(z)] \\ &+ G_1(0, 0)\sigma_2(\lambda_1 + \lambda_2)\sigma_1(\lambda_1 + \lambda_2(1 - z_2))[1 - \eta_2^{(1)}(0, z_2)][1 - \eta_1^{(1)}(z)]. \end{aligned} \tag{4.11}$$

Putting $z = (1, 0)$, $z = (0, 1)$, and $z = (0, 0)$ in (4.11), we obtain a set of three linear equations for the unknown constants $G_1(1, 0)$, $G_1(0, 1)$, and $G_1(0, 0)$.

*Case VI.* $n = 2$ *queues,* $\eta_1(z) = z_1$, $\eta_2(z) = 1$.
This covers the case $L_1 = 1$, $L_2 = \infty$ described in the previous section, i.e., one of the customer present at the start of a visit to $Q_1$ is served, customers arriving at $Q_1$ during the course of a visit are served with probability $p_1$, all the customers present at the start of a visit to $Q_2$ are served, customers arriving at $Q_2$ during the course of a visit are served with probability $p_2$, moreover assuming that there are an infinite number of servers and deterministic service times at $Q_2$, cf. Example 3.2.
Equation (4.3) reduces to

$$\begin{aligned} G_1(z_1, z_2) &= G_1(z_1, 1)\sigma_2(\gamma(z_1, 1))\sigma_1(\gamma(z))\eta_2^{(1)}(z)\frac{\eta_1^{(1)}(z)}{z_1} \\ &+ G_1(z_1, 0)\sigma_2(\gamma(z_1, 0))\sigma_1(\gamma(z))[1 - \eta_2^{(1)}(z)]\frac{\eta_1^{(1)}(z)}{z_1} \\ &+ G_1(0, 1)\sigma_2(\gamma(0, 1))\sigma_1(\gamma(0, z_2))\eta_2^{(1)}(0, z_2)[1 - \frac{\eta_1^{(1)}(z)}{z_1}] \\ &+ G_1(0, 0)\sigma_2(\gamma(0, 0))\sigma_1(\gamma(0, z_2))[1 - \eta_2^{(1)}(0, z_2)][1 - \frac{\eta_1^{(1)}(z)}{z_1}]. \end{aligned} \tag{4.12}$$

Setting $z_2 = 0$ and $z_2 = 1$ in (4.12) we find expressions for $G_1(z_1, 0)$ and $G_1(z_1, 1)$ containing the unknown constants $G_1(0, 0)$ and $G_1(0, 1)$. Putting $z_1 = 0$ in those expressions we obtain a pair of linear equations for these constants.

*Case VII. general* $n$, $\eta_i(z) = 1$, $i = 1, \ldots, n$.
Similar to Case V.

*Case VIII. general* $n$, $\eta_1(z) = z_1$, $\eta_i(z) = 1$, $i \neq 1$.
Similar to Case VI.

## 5 Concluding remarks and suggestions for further research

So far we focused on the joint queue length distribution at embedded epochs. In section 3 we obtained under Assumption 3.2 a set of $2n$ equations (3.2), (3.13) for the associated pgf's $F_i(z)$, $G_i(z)$, $i = 1, \ldots, n$. In section 4 we identified some cases in which these pgf's can actually be solved from these equations. These cases include several single-queue systems with a varying number of servers, two-queue two-server systems with exhaustive service and exponential service times, as well as infinite-server systems with an arbitrary number of queues, exhaustive or gated service, and deterministic service times.

To conclude, we now briefly discuss the derivation of the marginal queue length distribution at an arbitrary epoch from the joint queue length distribution at embedded epochs. Denote by $\mathbf{N}_i$ the queue length at $Q_i$ at an arbitrary epoch. As stated in the introduction, in isolation a particular queue in a polling system may be viewed as a single-queue system with service interruptions, the intervisit periods constituting the service interruptions. In section 2 we showed how in such a system with service interruptions and exponential service times, the queue length distribution at an arbitrary epoch may be expressed into the queue length distribution at the beginning and the end of a service interruption. In case the assumptions of section 2 are satisfied, one may thus obtain the marginal queue length distribution at $Q_i$ from the queue length distribution at the beginning and the end of a visit to $Q_i$, given by $\mathrm{E}(z^{\mathbf{X}_{ii}}) = F_i(1, \ldots, 1, z, 1, \ldots, 1)$ and $\mathrm{E}(z^{\mathbf{Y}_{ii}}) = G_i(1, \ldots, 1, z, 1, \ldots, 1)$, respectively, with $z$ as $i$-th argument. Consider e.g. the two-queue two-server system with exhaustive service and exponential service times, for which we obtained $F_i(z)$ and $G_i(z)$ in Case IV of the previous section. For such a system, using Lemma 2.1 and Lemma 2.2,

$$\mathrm{E}(z^{\mathbf{N}_i}) = \left[\frac{2}{2-\rho_i} + \frac{\rho_i}{2-\rho_i}\Pr\{\mathbf{N}_{A,i} = 0\}\right]^{-1} \left[\frac{2}{2-\rho_i z}\mathrm{E}(z^{\mathbf{N}_{A,i}}) + \frac{\rho_i z}{2-\rho_i z}\Pr\{\mathbf{N}_{A,i} = 0\}\right],$$

with

$$\mathrm{E}(z^{\mathbf{N}_{A,i}}) = \frac{1 - \mathrm{E}(z^{\mathbf{X}_{ii}})}{(1-z)\mathrm{E}\mathbf{X}_{ii}}.$$

In section 2 we also showed how subsequently the waiting time distribution may be related to the marginal queue length distribution by using Lemma 2.3.

In case the assumptions of section 2 are not satisfied, one may quite often still obtain the marginal queue length distribution from the joint queue length distribution at the beginning and the end of a visit by developing ad hoc methods. We do however not pursue the matter any further, leaving it as an interesting topic for further research.

**Acknowledgement** The author is grateful to O.J. Boxma for several valuable discussions and useful suggestions.

## References

[1] Ajmone Marsan, M., Donatelli, S., Neri, F. (1990). GSPN models of Markovian multiserver multiqueue systems. *Perf. Eval.* **11**, 227-240.

[2] Ajmone Marsan, M., Donatelli, S., Neri, F. (1991). Multiserver multiqueue systems with limited service and zero walk time. In: *Proc. INFOCOM '91*, 1178-1188.

[3] Ajmone Marsan, M., De Moraes, L.F., Donatelli, S., Neri, F. (1990). Analysis of symmetric nonexhaustive polling with multiple servers. In: *Proc. INFOCOM '90*, 284-295.

[4] Ajmone Marsan, M., De Moraes, L.F., Donatelli, S., Neri, F. (1992). Cycles and waiting times in symmetric exhaustive and gated multiserver multiqueue systems. In: *Proc. INFOCOM '92*, 2315-2324.

[5] Arem, B. van (1990). Queueing Models for Slotted Transmission Systems. *Ph.D. Thesis, Twente University, Enschede.*

[6] Bhuyan, L.N., Ghosal, D., Yang, Q. (1989). Approximate analysis of single and multiple ring networks. *IEEE Trans. Comp.* **38**, 1027-1040.

[7] Bisdikian, C. (1993). The random N-policy. Report IBM Yorktown Heights.

[8] Boxma, O.J. (1989). Workloads and waiting times in single-server queues with multiple customer classes. *Queueing Systems* **5**, 185-214.

[9] Boxma, O.J., Groenendijk, W.P. (1988). Two queues with alternating service and switching times. In: *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen*, eds. O.J. Boxma and R. Syski (North-Holland, Amsterdam), 261-282.

[10] Browne, S., Coffman, E.G. jr., Gilbert, E.N., Wright, P.E.W. (1992). Gated, exhaustive, parallel service. *Prob. Eng. Inf. Sc.* **6**, 217-239.

[11] Browne, S., Kella, O. (1992). Parallel service with vacations. Report Columbia University.

[12] Browne, S., Weiss, G. (1992). Dynamic priority rules when polling with multiple parallel servers. *Oper. Res. Lett.* **12**, 129-137.

[13] Fuhrmann, S.W. (1981). Performance analysis of a class of cyclic schedules, Bell Laboratories technical memorandum 81-59531-1.

[14] Fuhrmann, S.W., Cooper, R.B. (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper. Res.* **33**, 1117-1129.

[15] Gamse, B., Newell, G.F. (1982). An analysis of elevator operation in moderate height buildings - II. Multiple elevators. *Transp. Res., B* **16**, 321-335.

[16] Kamal, A.E., Hamacher, V.C. (1989). Approximate analysis of non-exhaustive multiserver polling systems with applications to local area networks. In: *Comp. Netw. ISDN Syst.* **17**, 15-27.

[17] Kao, E.P.C., Narayanan, K.S. (1991). Analyses of an $M/M/N$ queue with servers' vacations. *EJOR* **54**, 256-266.

[18] Karmarkar, V.V., Kuhl, J.G. (1989). An integrated approach to distributed demand assignment in multiple-bus local networks. *IEEE Trans. Comp.* **38**, 679-695.

[19] Levy, Y., Yechiali, U. (1976). An $M/M/s$ queue with servers' vacations. *INFOR* **14**, 153-163.

[20] Loucks, W.M., Hamacher, V.C., Preiss, B.R., Wong, L. (1985). Short-packet transfer performance in local area ring networks. *IEEE Trans. Comp.* **34**, 1006-1014.

[21] Medhi, J. (1991). *Stochastic Models in Queueing Theory.* (Academic Press, San Diego).

[22] Mitrany, I.L., Avi-Itzhak, B. (1968). A many-server queue with server interruptions. *Oper. Res.* **16**, 628-638.

[23] Morris, R.J.T., Wang, Y.T. (1984). Some results for multi-queue systems with multiple cyclic servers. In: *Performance of Computer-Communication Systems*, eds. W. Bux and H. Rudin (North-Holland, Amsterdam), 245-258.

[24] Neuts, M.F., Lucantoni, D.M. (1979). A Markovian queue with $N$ servers subject to breakdowns and repairs. *Mgmt. Sc.* **25**, 849-861.

[25] Raith, T. (1985). Performance analysis of multibus interconnection networks in distributed systems. In: *Proc. ITC-11*, 662-668.

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

[27] Stadje, W. (1985). The busy period of the queueing system $M/G/\infty$. *J. Appl. Prob.* **22**, 697-704.

[28] Titchmarsh, E.C. (1939). *The Theory of Functions.* (Oxford University Press, London, 2nd ed.).

[29] Yang, Q., Ghosal, D., Bhuyan, L.N. (1986). Performance analysis of multiple token ring and multiple slotted ring networks. In: *Proc. 1986 Comp. Netw. Symp.*, 79-86.

[30] Zafirovic-Vukotic, M., Niemegeers, I.G., Valk, D.S. (1988). Performance modelling of slotted ring protocols in HSLAN's. *IEEE J. Sel. Areas Comm.* **6**, 1001-1024.

## Appendices

### A Proof of Lemma 2.1

**Lemma 2.1**

$$\mathrm{E}(z^{\mathbf{N}}) = \gamma \left[ \sum_{l=0}^{m-2} \frac{\mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}}) \Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}} + \frac{m}{m-\rho z} \sum_{l=m-1}^{\infty} z^l \Pr\{\mathbf{N}_A = l\} \right],$$

with

$$\gamma = \left[ \sum_{l=0}^{m-2} \frac{\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}} + \frac{m}{m-\rho} \sum_{l=m-1}^{\infty} \Pr\{\mathbf{N}_A = l\} \right]^{-1}.$$

**Proof**

Define a vacation customer to be a customer arriving in a non-serving interval. Consider now a vacation customer $C$ arriving at some time $u$. Suppose that $C$ sees $l$ customers upon arrival; so the queue length just after $u$ equals $l+1$, $l \geq 0$. Let $\mathbf{T}$ be the first epoch in a serving interval after $u$ at which the queue length reaches the level $l+1$ again. Let $\mathbf{U}$ be the first epoch after $u$ at which the queue length drops to the level $l$. Suppose that the interval $[\mathbf{T}, \mathbf{U}]$ contains $\mathbf{K}$ distinct non-serving intervals starting at the consecutive epochs $\mathbf{U}_1, \ldots, \mathbf{U}_\mathbf{K}$, $\mathbf{K} \geq 0$. Let $\mathbf{N}_k$ be the queue length just after the epoch $\mathbf{U}_k$. Let $\mathbf{T}_k$ be the first epoch in a serving interval after $\mathbf{U}_k$ at which the queue length reaches the level $\mathbf{N}_k$ again. The interval $[\mathbf{T}, \mathbf{U}]$, exclusive of the intervals $[\mathbf{U}_1, \mathbf{T}_1], \ldots, [\mathbf{U}_\mathbf{K}, \mathbf{T}_\mathbf{K}]$, is called a 1-busy period at level $l$. Note that we have thus established a 1-1 correspondence between 1-busy periods at level $l$ and vacation customers that see $l$ customers upon arrival. (For $m = 1$ one can establish the 1-1 correspondence in an elegant way by choosing the order of service to be non-preemptive LCFS, cf. Fuhrmann & Cooper [14]; the vacation customer is then the 'ancestor' of the customers served in the 1-busy period. For $m > 1$ one can not establish the 1-1 correspondence in such an elegant way, as the customers then do not necessarily leave in order of service.) The notion of a 1-busy period is illustrated in Figure I, with $\mathbf{N}(t)$ denoting the queue length at time $t$. Parallel to the time axis the non-serving intervals are indicated by dotted lines. The serving intervals constituting a 1-busy period at level $l = 1$ are indicated by bold lines.

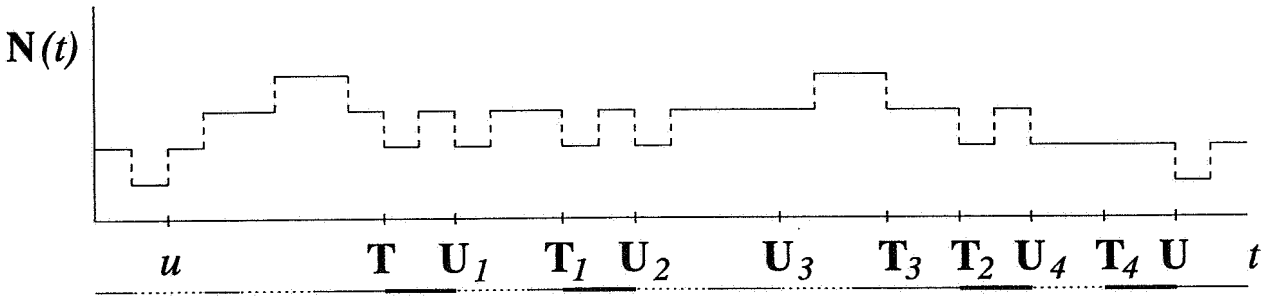


Figure I. A 1-busy period at level $l = 1$.

Consider now an arbitrary tagged customer as it departs from the system.
Denote by $\mathbf{N}_D$ the number of customers that the tagged customer leaves behind. By virtue of the PASTA property and an up- & down crossing argument, $\mathbf{N}_D$ has the same distribution as $\mathbf{N}$. Denote by $\mathbf{L}_D$ the level of the 1-busy period in which the tagged customer is served. (Note here that the 1-busy periods together constitute a partitioning of the serving intervals.)

$$\mathrm{E}(z^{\mathbf{N}}) = \mathrm{E}(z^{\mathbf{N}_D}) = \sum_{l=0}^{\infty} \mathrm{E}(z^{\mathbf{N}_D} \mid \mathbf{L}_D = l)\mathrm{Pr}\{\mathbf{L}_D = l\}. \tag{A.1}$$

Define a 1-busy period at level $l$ in the corresponding $M/M/m$ queue to be a period ranging from an epoch when the queue length jumps to the level $l+1$ to an epoch when the queue length drops to the level $l$. Because of the memoryless property of the exponential service time distribution, a 1-busy period at level $l$ in the queue with service interruptions is stochastically indistinguishable from a 1-busy period at level $l$ in the corresponding $M/M/m$ queue. So,

given that $\mathbf{L}_D = l$, $\mathbf{N}_D$ has the same distribution as the number of customers that an arbitrary customer leaves behind as it departs from the corresponding $M/M/m$ queue in a 1-busy period at level $l$. By virtue of the (conditional) PASTA property and an up- & down crossing argument, this number has again the same distribution as $\mathbf{N}^{(l)}_{M/M/m}$, the queue length at an arbitrary epoch in the corresponding $M/M/m$ queue given that the queue length is at least $l$.

$$\mathrm{E}(z^{\mathbf{N}_D} \mid \mathbf{L}_D = l) = \mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}}). \tag{A.2}$$

Denote by $\mathbf{L}$ the level of an arbitrary 1-busy period. Remember that we have established a 1-1 correspondence between 1-busy periods at level $l$ and vacation customers that see $l$ customers upon arrival. So $\mathbf{L}$ has the same distribution as the number of customers seen by an arbitrary arriving vacation customer. Because of the PASTA property, this number has again the same distribution as $\mathbf{N}_A$. Denote by $\mathbf{M}_l$ the number of customers served in a 1-busy period at level $l$. Then

$$\Pr\{\mathbf{L}_D = l\} = \frac{\Pr\{\mathbf{L} = l\}\mathrm{E}\mathbf{M}_l}{\sum_{k=0}^{\infty} \Pr\{\mathbf{L} = k\}\mathrm{E}\mathbf{M}_k} = \frac{\Pr\{\mathbf{N}_A = l\}\mathrm{E}\mathbf{M}_l}{\sum_{k=0}^{\infty} \Pr\{\mathbf{N}_A = k\}\mathrm{E}\mathbf{M}_k}. \tag{A.3}$$

In a 1-busy period at level $l$ exactly 1 customer is served that leaves behind $l$ customers as it departs from the system. So $\mathrm{E}\mathbf{M}_l$ equals the reciprocal of the probability that an arbitrary customer leaves behind $l$ customers as it departs from the system in a 1-busy period at level $l$:

$$\mathrm{E}\mathbf{M}_l = \frac{1}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}. \tag{A.4}$$

Summarizing,

$$\mathrm{E}(z^{\mathbf{N}}) = \gamma \left[ \sum_{l=0}^{\infty} \frac{\mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}})\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}} \right], \tag{A.5}$$

with

$$\gamma = \left[ \sum_{k=0}^{\infty} \frac{\Pr\{\mathbf{N}_A = k\}}{\Pr\{\mathbf{N}^{(k)}_{M/M/m} = k\}} \right]^{-1}. \tag{A.6}$$

Substituting (2.2) into (A.5) and (A.6) completes the proof.

□

## B Proof of Lemma 4.1

**Lemma 4.1**

If $\eta'(1) < 1$ then

*i.* $\lim_{K\to\infty} e^{(K+1)}(z) = 1$ for all $z$ with $|z| \le 1$;

*ii.* $\prod_{k=0}^{\infty} f(e^{(k)}(z))$ converges for all $z$ with $|z| \le 1$;

*iii.* $\sum_{k=0}^{\infty} g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z))$ converges for all $z$ with $|z| \le 1$.

**Proof**

*Proof of i.* Since $e(z) = \eta(z)$ is a pgf,

$$|1 - e(z)| \le \eta'(1)\,|1 - z|\,.$$

By induction,

$$|1 - e^{(k)}(z)| \le (\eta'(1))^k\,|1 - z|, \qquad k \ge 0. \tag{B.1}$$

So $\lim_{K\to\infty} e^{(K+1)}(z) = 1$.

*Proof of ii.* According to the theory of infinite products, cf. Titchmarsh [28] p. 18, $\prod_{k=0}^{\infty} f(e^{(k)}(z))$ converges iff $\sum_{k=0}^{\infty}[1 - f(e^{(k)}(z))]$ converges.

Let $\Gamma(z)$ be the straight contour in the complex plane from $z$ to 1.

According to the theory of complex functions,

$$|1 - f(z)| = |f(1) - f(z)| = |\int_{u\in\Gamma(z)} df(u)| \le M(z)\,|1 - z|, \tag{B.2}$$

with

$$M(z) = \max_{u\in\Gamma(z)} |\frac{df(u)}{du}| < \infty,$$

as $f(u)$ is continuously-differentiable on $|u| \le 1$.

Using (B.1), (B.2),

$$\sum_{k=0}^{\infty} |1 - f(e^{(k)}(z))| \le \sum_{k=0}^{\infty} M(z)(\eta'(1))^k\,|1 - z| = \frac{M(z)}{1 - \eta'(1)}\,|1 - z| < \infty.$$

So $\prod_{k=0}^{\infty} f(e^{(k)}(z))$ converges.

*Proof of iii.* Note that $\sum_{k=0}^{\infty} |g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z))| \leq K \sum_{k=0}^{\infty} |g(e^{(k)}(z)|$ with $K = \max_{k\geq 0} |\prod_{l=0}^{k-1} f(e^{(l)}(z))|$.
As $\prod_{k=0}^{\infty} f(e^{(k)}(z))$ converges, $\max_{k\geq 0} |\prod_{l=0}^{k-1} f(e^{(l)}(z))| < \infty$. So to prove that $\sum_{k=0}^{\infty} g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z))$ converges, it suffices to prove that $\sum_{k=0}^{\infty} g(e^{(k)}(z))$ converges.

Let $\Gamma(z)$ be the straight contour in the complex plane from $z$ to 1.
Similarly to (B.2), noting that $g(1) = 0$,

$$|g(z)| \leq N(z)\,|1-z|, \tag{B.3}$$

with

$$N(z) = \max_{u\in\Gamma(z)} |\frac{dg(u)}{du}| < \infty,$$

as $g(u)$ is also continuously-differentiable on $|u| \leq 1$.
Using (B.1), (B.3),

$$\sum_{k=0}^{\infty} |g(e^{(k)}(z))| \leq \sum_{k=0}^{\infty} N(z)(\eta'(1))^k\,|1-z| = \frac{N(z)}{1-\eta'(1)}\,|1-z| < \infty.$$

So $\sum_{k=0}^{\infty} g(e^{(k)}(z)) \prod_{l=0}^{k-1} f(e^{(l)}(z))$ converges.

□