Queueing Systems 20 (1995) 369–393

Polling systems with multiple coupled servers

S.C. Borst

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Received 12 July 1994; revised 5 February 1995

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.

Keywords: 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 by multiple servers in a cyclic manner. 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.

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. [22]. 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.

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 packet, circulating at constant speed. Occupying a slot corresponds to utilizing a server. Another medium access variant that may lead to multiple-server polling is the token ring, i.e., there are

multiple rings, each with a token circulating on it, representing the right of transmission on that particular ring. Holding a token corresponds to utilizing a server.

Multiple-server polling systems have received remarkably little attention in the vast literature on polling systems (see Takagi [26] for a comprehensive survey). One of the first studies is Morris and Wang [22] in which the servers are assumed to be independent, i.e., to visit the queues independently of each other, each server according to its own 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 and Wang is the tendency for the servers to cluster if they follow identical routes, especially in heavy traffic. 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.

Browne and Weiss [11] 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. [9] derive the mean waiting time for a completely symmetric two-queue system with an infinite number of coupled servers and deterministic service times. Browne and Kella [10] 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.

Levy and Yechiali [19] and Kao and 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 and Avi-Itzhak [21] and Neuts and Lucantoni [23] 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 refs. [6, 12, 16, 18, 24, 29, 30] mean response time approximations are developed to analyze the performance of LAN's with multiple-token rings. Mean response time approximation oriented to LAN's with a multiple-slotted ring are contained in refs. [5, 6, 20, 29, 31]. Ajmone Marsan et al. [2–4] 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 [9, 10] and the single-queue studies [17, 19, 21, 23]) presents any exact results, apart from some mean-value results for global performance

measures like cycle times. Indeed, in a Markov description the state of the system has to be represented by a vector containing the number of customers at each of the queues, the position of each of the servers, as well as the remaining service or switch-over time for each of the servers. In general, the state evolution process is prohibitively complex, involving an intractable infinite set of difference-differential equations.

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 threefold. 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 and Cooper [4] 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 and Cooper, Boxma [7] proves a similar decomposition of the amount of work in the system. Browne and Kella [10] analyze the queue length distribution in an $M/G/\infty$ queue with vacations. They observe that for deterministic service times a Fuhrmann and 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 λ be the arrival rate and let μ 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, $l \geqslant 0$. The "corresponding" $M/M/m$ queue is an ordinary $M/M/m$ queue with arrival rate λ and service rate μ.
For $l \leqslant m-1$

$$\Pr\{\mathbf{N}^{(l)}_{M/M/m} = k\} = \begin{cases} \gamma^{(l)} \dfrac{\rho^k}{k!}, & l \leqslant k \leqslant m-1, \\ \gamma^{(l)} \dfrac{\rho^m}{m!} \left(\dfrac{\rho}{m}\right)^{k-m}, & k \geqslant m, \end{cases} \tag{2.1}$$

so that

$$\mathrm{E}\left(z^{\mathbf{N}^{(l)}_{M/M/m}}\right) = \gamma^{(l)} \left[\sum_{k=l}^{m-1} z^k \frac{\rho^k}{k!} + z^m \frac{\rho^m}{m!} \frac{m}{m-\rho z}\right], \tag{2.2}$$

with

$$\gamma^{(l)} = \left[\sum_{k=l}^{m-1} \frac{\rho^k}{k!} + \frac{\rho^m}{m!} \frac{m}{m-\rho}\right]^{-1}. \tag{2.3}$$

For $l \geqslant m-1$,

$$\Pr\{\mathbf{N}^{(l)}_{M/M/m} = k\} = \frac{m-\rho}{m} \left(\frac{\rho}{m}\right)^{k-l}, \quad k \geqslant l, \tag{2.4}$$

so that

$$\mathrm{E}\left(z^{\mathbf{N}^{(l)}_{M/M/m}}\right) = z^l \frac{m-\rho}{m-\rho z}. \tag{2.5}$$

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

LEMMA 2.1

$$\mathrm{E}\left(z^{\mathbf{N}}\right) = \left[\sum_{l=0}^{\infty} \frac{\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}\right]^{-1} \left[\sum_{l=0}^{\infty} \frac{\Pr\{\mathbf{N}_A = l\} \mathrm{E}\left(z^{\mathbf{N}^{(l)}_{M/M/m}}\right)}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}\right]. \tag{2.6}$$

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 $l+1$, $l \geqslant 0$. 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 $\mathbf{U}_1, \ldots, \mathbf{U}_{\mathbf{K}}, \mathbf{K} \geqslant 0$. 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 and Cooper [14]; the vacation customer is then the "ancestor" of the customers served in the 1-busy period. For $m > 1$ one cannot 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 fig. 1, 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.

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- and downcrossing 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.)

$$\mathrm{E}(z^{\mathbf{N}}) = \mathrm{E}(z^{\mathbf{N}_D}) = \sum_{l=0}^{\infty} \mathrm{E}(z^{\mathbf{N}_D} | \mathbf{L}_D = l) \Pr\{\mathbf{L}_D = l\}. \tag{2.7}$$

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

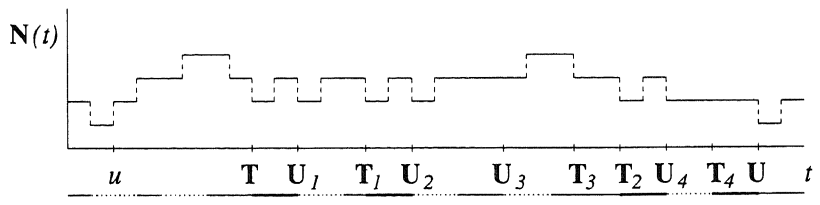

Fig. 1. A 1-busy period at level $l = 1$.

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- and downcrossing 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.

$$\mathrm{E}(z^{\mathbf{N}_D}|\mathbf{L}_D = l) = \mathrm{E}(z^{\mathbf{N}^{(l)}_{M/M/m}}). \tag{2.8}$$

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 N_A. Denote by $\mathbf{M}_l$ the number of customers served in a 1-busy period at level l. Then

$$\Pr\{\mathbf{L}_D = l\} = \frac{\Pr\{\mathbf{L} = l\}\mathrm{E}\mathbf{M}_l}{\sum_{k=0}^{\infty}\Pr\{\mathbf{L} = k\}\mathrm{E}\mathbf{M}_k} = \frac{\Pr\{\mathbf{N}_A = l\}\mathrm{E}\mathbf{M}_l}{\sum_{k=0}^{\infty}\Pr\{\mathbf{N}_A = k\}\mathrm{E}\mathbf{M}_k}. \tag{2.9}$$

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:

$$\mathrm{E}\mathbf{M}_l = \frac{1}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}. \tag{2.10}$$

Substituting (2.8)–(2.10) into (2.7) completes the proof. □

Substituting (2.1)–(2.5) into (2.6) yields

$$\mathrm{E}(z^{\mathbf{N}}) = \gamma\left[\sum_{l=0}^{m-2}\Pr\{\mathbf{N}_A = l\}\frac{l!}{\rho^l}\left[\sum_{k=l}^{m-1} z^k\frac{\rho^k}{k!} + z^m\frac{\rho^m}{m!}\frac{m}{m-\rho z}\right] + \frac{m}{m-\rho z}\sum_{l=m-1}^{\infty}\Pr\{\mathbf{N}_A = l\}z^l\right], \tag{2.11}$$

with

$$\gamma = \left[\sum_{l=0}^{m-2}\Pr\{\mathbf{N}_A = l\}\frac{l!}{\rho^l}\left[\sum_{k=l}^{m-1}\frac{\rho^k}{k!} + \frac{\rho^m}{m!}\frac{m}{m-\rho}\right] + \frac{m}{m-\rho}\sum_{l=m-1}^{\infty}\Pr\{\mathbf{N}_A = l\}\right]^{-1}. \tag{2.12}$$

Remark 2.1

For $\Pr\{\mathbf{N}_A = 0\} = 1$, i.e., in a non-serving interval there are never any customers present, (2.6), (2.11) and (2.12) reduce to

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

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 $m = 1$, (2.11) and (2.12) 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 $m = \infty$, (2.11) and (2.12) reduce to

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

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

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

which may be used to recover a result first obtained in Browne and Kella [10].

The above results imply that to find the distribution of $\mathbf{N}$, it suffices to find the distribution of $\mathbf{N}_A$. 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 $\mathbf{N}_B^{(k)}$, $\mathbf{N}_C^{(k)}$. The following equation (see Wolff [28] Section 10.5 for a similar relationship) relates the distribution of $\mathbf{N}_A$ to the distribution of $\mathbf{N}_B$ and $\mathbf{N}_C$.

$$\Pr\{\mathbf{N}_A = l\} = \frac{\Pr\{\mathbf{N}_B \leqslant l\} - \Pr\{\mathbf{N}_C \leqslant l\}}{\mathrm{E}\mathbf{N}_C - \mathrm{E}\mathbf{N}_B}. \tag{2.13}$$

Written in terms of pgf's

$$E(z^{N_A}) = \frac{E(z^{N_B}) - E(z^{N_C})}{(1-z)(EN_C - EN_B)}. \tag{2.14}$$

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 $\mathbf{W}$ and $\mathbf{R}$ the waiting and the sojourn time, respectively, of an arbitrary customer. Denote by $\mathbf{L}$ the number of waiting customers at an arbitrary epoch. The familiar relationship $E(z^{N}) = E(e^{-\lambda(1-z)\mathbf{R}})$ 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 $E(z^{L}) = E(e^{-\lambda(1-z)\mathbf{W}})$. What thus remains to be done, is to relate the distribution of $\mathbf{L}$ to the distribution of $\mathbf{N}$.

LEMMA 2.2

$$E(z^{L}) = E(z^{N}) + (1-p_A)\left([z^{-m} - 1]E(z^{N_E}) + \sum_{k=0}^{m-1} [1 - z^{k-m}]\Pr\{N_E = k\} \right), \tag{2.15}$$

with

$$\Pr\{N_E = k\} = \frac{\Pr\{N = k\} - p_A \Pr\{N_A = k\}}{1 - p_A}, \tag{2.16}$$

$$E(z^{N_E}) = \frac{E(z^{N}) - p_A E(z^{N_A})}{1 - p_A}, \tag{2.17}$$

$$p_A = \left[\sum_{l=0}^{\infty} \frac{\Pr\{N_A = l\}}{\Pr\{N_{M/M/m}^{(l)} = l\}} \right]^{-1}. \tag{2.18}$$

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

$$\mathrm{E}(z^{\mathbf{N}}) = \mathrm{E}(z^{\mathbf{N}_A})p_A + \mathrm{E}(z^{\mathbf{N}_E})(1-p_A), \tag{2.19}$$

$$\mathrm{E}(z^{\mathbf{L}}) = \mathrm{E}(z^{\mathbf{N}_A})p_A + \mathrm{E}(z^{[\mathbf{N}_E-m]^+})(1-p_A), \tag{2.20}$$

with $[x]^+ = \max(0,x)$.

Comparing (2.19), (2.20), using that

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

yields (2.15).

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

$$p_A = \frac{1}{\mathrm{E}\mathbf{M}}. \tag{2.21}$$

From the proof of lemma 2.1

$$\mathrm{E}\mathbf{M} = \sum_{l=0}^{\infty} \frac{\Pr\{\mathbf{N}_A = l\}}{\Pr\{\mathbf{N}^{(l)}_{M/M/m} = l\}}. \tag{2.22}$$

Substituting (2.22) into (2.21) 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 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 λ_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 Laplace-Stieltjes Transform (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 $(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in})$ and $(\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in})$ the joint queue length vector at, respectively, the beginning and the end of a visit to $Q_i, i = 1, \ldots, n$. 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 $z = (z_1, \ldots, z_n), |z_h| \leqslant 1, h = 1, \ldots, n, i = 1, \ldots, n$.

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

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

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

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.

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

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

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 [25] 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 $\eta_i(z) = \mathrm{E}(z_1^{\mathbf{K}_{i1}} \cdots z_n^{\mathbf{K}_{in}})$.

Formally,

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

Substituting (3.4) into (3.3),

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

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,

$$\mathrm{E}\big(z_1^{\mathbf{Y}_{i1}} \cdots z_n^{\mathbf{Y}_{in}} | (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in}) = (l_1, \ldots, l_n)\big) = z_1^{l_1} \cdots z_{i-1}^{l_{i-1}} \eta_i^{(l_i)}(z) z_{i+1}^{l_{i+1}} \cdots z_n^{l_n}, \tag{3.6}$$

with

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

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

Substituting (3.6), (3.7), (3.8) into (3.3),

$$G_i(z) = F_i(z_1, \ldots, z_{i-1}, \eta_i(z), z_{i+1}, \ldots, z_n)\frac{\eta_i^{(1)}(z)}{\eta_i(z)} + F_i(z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_n)\left[1 - \frac{\eta_i^{(1)}(z)}{\eta_i(z)}\right]. \tag{3.9}$$

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

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 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. Keeping this in mind, we consider a class of service disciplines that are parametrized by two vectors $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$ with the following interpretation. If there are any customers present at Q_i at the start of a visit, then one of them is always served, while the others are served with probability q_i. Customers arriving at Q_i during the course of a visit

are served with probability p_i. The case $q_i = 0$ contains both the semi-exhaustive service discipline ($p_i = 1$) and the 1-limited service discipline ($p_i = 0$). The case $q_i = 1$ 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 $T_i^{(k)}$ be the length of a busy period starting with k customers present in an ordinary $M/G/m$ queue with arrival rate κ_i and service time distribution $B_i(\cdot)$. Let $\tau_i^{(k)}(\omega) = E(e^{-\omega T_i^{(k)}})$ for $\mathrm{Re}\,\omega \geqslant 0$.

Define

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

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

For the above-defined class of service disciplines,

$$\eta_i^{(l_i)}(z) = 1, \quad l_i = 0, \tag{3.14}$$

$$\eta_i^{(l_i)}(z) = \sum_{k_i=1}^{l_i} \binom{l_i - 1}{k_i - 1} q_i^{k_i - 1}(1 - q_i)^{l_i - k_i} \tau_i^{(k_i)}(\alpha_i(z)) z_i^{l_i - k_i}, \quad l_i > 0, \tag{3.15}$$

with the interpretation of $\eta_i^{(l_i)}(z)$ as in (3.7), (3.8). For $q_i = 0$, (3.14), (3.15) satisfy (3.7), (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

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

for some LST $\tau_i(\cdot)$, then (3.14), (3.15) satisfy (3.7), (3.8) also for $q_i > 0$ with $\eta_i(z) = q_i\tau_i(\alpha_i(z)) + (1 - q_i)z_i, \eta_i^{(1)}(z) = \tau_i^{(1)}(\alpha_i(z))$.

Two examples where $\tau_i^{(k_i)}(\cdot)$ is of the above form are (i) the case of $m = 2$ servers and exponential service times (which is one of the examples in Browne and Weiss [11]), and (ii) the case of $m = \infty$ servers and deterministic service times (which is one of the examples in Browne et al. [9] and Browne and Kella [10]).

4. The joint queue length distribution II

In the previous section we obtained under assumption 3.2 a set of $2n$ eqs. (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) yields

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

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),

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

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),

$$\begin{aligned} G_1(z_1, z_2) = {} & G_1(\eta_1(z), \eta_2(\eta_1(z), z_2))\sigma_2\{\gamma(\eta_1(z), \eta_2(\eta_1(z), z_2))\}\sigma_1\{\gamma(\eta_1(z), z_2)\} \\ & \times \frac{\eta_2^{(1)}(\eta_1(z), z_2)}{\eta_2(\eta_1(z), z_2)}\frac{\eta_1^{(1)}(z)}{\eta_1(z)} + G_1(\eta_1(z), 0)\sigma_2\{\gamma(\eta_1(z), 0)\}\sigma_1\{\gamma(\eta_1(z), z_2)\} \\ & \times \left[1 - \frac{\eta_2^{(1)}(\eta_1(z), z_2)}{\eta_2(\eta_1(z), z_2)}\right]\frac{\eta_1^{(1)}(z)}{\eta_1(z)} \\ & + G_1(0, \eta_2(0, z_2))\sigma_2\{\gamma(0, \eta_2(0, z_2))\}\sigma_1\{\gamma(0, z_2)\} \\ & \times \frac{\eta_2^{(1)}(0, z_2)}{\eta_2(0, z_2)}\left[1 - \frac{\eta_1^{(1)}(z)}{\eta_1(z)}\right] + G_1(0, 0)\sigma_2\{\gamma(0, 0)\}\sigma_1\{\gamma(0, z_2)\} \\ & \times \left[1 - \frac{\eta_2^{(1)}(0, z_2)}{\eta_2(0, z_2)}\right]\left[1 - \frac{\eta_1^{(1)}(z)}{\eta_1(z)}\right], \end{aligned} \tag{4.3}$$

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

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 cannot 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 $q = 0$ 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 $(\eta^{(1)})'(1) = (d\eta^{(1)}(z)/dz)|_{z=1}$. Apparently the stability condition is $\lambda s + (\eta^{(1)})'(1) < 1$. Note that $\lambda s + (\eta^{(1)})'(1)$ is the mean increase in 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 $q > 0$ described in the previous section, i.e., one of the customers present at the start of a visit is always served, the others are served with probability q, 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 or an infinite number of servers and deterministic service times. 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

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

for $|z| \leqslant 1$.

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 $|z| \leqslant 1$;

(ii) $\prod_{k=0}^{\infty} f\left(e^{(k)}(z)\right)$ converges for all z with $|z| \leqslant 1$;

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

Proof

Proof of i

Since $e(z) = \eta(z)$ is a pgf,

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

By induction,

$$|1 - e^{(k)}(z)| \leqslant (\eta'(1))^k |1 - z|, \quad k \geqslant 0. \tag{4.7}$$

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

Proof of ii

According to the theory of infinite products, cf. Titchmarsh [27, p. 18], $\prod_{k=0}^{\infty} f\left(e^{(k)}(z)\right)$ converges iff $\sum_{k=0}^{\infty} \left[1 - f\left(e^{(k)}(z)\right)\right]$ converges.

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

According to the theory of complex functions,

$$|1 - f(z)| = |f(1) - f(z)| = \left| \int_{u\in\Gamma(z)} df(u) \right| \leqslant M(z)|1 - z|, \tag{4.8}$$

with

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

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

Using (4.7), (4.8),

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

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

Proof of iii

Note that

$$\sum_{k=0}^{\infty} \left| g\left(e^{(k)}(z)\right) \prod_{l=0}^{k-1} f\left(e^{(l)}(z)\right) \right| \leqslant K \sum_{k=0}^{\infty} \left| g\left(e^{(k)}(z)\right) \right|$$

with

$$K = \max_{k \geqslant 0} \left| \prod_{l=0}^{k-1} f\left(e^{(l)}(z)\right) \right|.$$

As $\prod_{k=0}^{\infty} f\left(e^{(k)}(z)\right)$ converges,

$$\max_{k \geqslant 0} \left| \prod_{l=0}^{k-1} f\left(e^{(l)}(z)\right) \right| < \infty.$$

So to prove that $\sum_{k=0}^{\infty} g\left(e^{(k)}(z)\right) \prod_{l=0}^{k-1} f\left(e^{(l)}(z)\right)$ converges, it suffices to prove that $\sum_{k=0}^{\infty} g\left(e^{(k)}(z)\right)$ converges.

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

Similarly to (4.8), noting that $g(1) = 0$,

$$|g(z)| \leqslant N(z)|1 - z|, \tag{4.9}$$

with

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

as $g(u)$ is also continuously-differentiable on $|u| \leqslant 1$.

Using (4.7), (4.9),

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

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

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 the case of two servers and exponential service times,

$$\eta'(1) = \frac{(1-p)\rho}{2 - p\rho} < 1$$

iff $\rho < 2$, irrespective of p. In the case of an infinite number of servers and deterministic service times, $\eta'(1) = 0$, also irrespective of p. If $\eta'(1) < 1$ then, letting $K \to \infty$ in (4.6),

$$G(z) = \prod_{k=0}^{\infty} f\left(e^{(k)}(z)\right) + G(0) \sum_{k=0}^{\infty} g\left(e^{(k)}(z)\right) \prod_{l=0}^{k-1} f\left(e^{(l)}(z)\right). \tag{4.10}$$

Putting $z = 0$ in (4.10),

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

Case III. $n = 2$ queues, $\eta_i(z) = z_i, i = 1, 2$

This covers the case $q_i = 0$ 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

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

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 and Groenendijk [8].

Case IV. $n = 2$ queues, $\eta_i(z), \eta_i^{(1)}(z)$ do not depend on $z_i, i = 1, 2$

This occurs in the case of two servers, exponential service times, and 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 eq. (4.3) may be replaced by

$$H_1(z_2) = H_1(e_1(z_2))f_1(z_2) + H_1(\eta_2(0))g_1(z_2) + H_1(0)h_1(z_2), \tag{4.11}$$

with

$$e_1(z_2) = \eta_2(\eta_1(z), z_2);$$

$$f_1(z_2) = \sigma_2\{\gamma(\eta_1(z), \eta_2(\eta_1(z), z_2))\}\sigma_1\{\gamma(\eta_1(z), z_2)\}\frac{\eta_2^{(1)}(\eta_1(z), z_2)}{\eta_2(\eta_1(z), z_2)}\frac{\eta_1^{(1)}(z)}{\eta_1(z)};$$

$$g_1(z_2) = \sigma_2\{\gamma(0, \eta_2(0, z_2))\}\sigma_1\{\gamma(0, z_2)\}\frac{\eta_2^{(1)}(0, z_2)}{\eta_2(0, z_2)}\left[1 - \frac{\eta_1^{(1)}(z)}{\eta_1(z)}\right];$$

$$h_1(z_2) = \sigma_2\{\gamma(\eta_1(z),0)\}\sigma_1\{\gamma(\eta_1(z),z_2)\}\left[1 - \frac{\eta_2^{(1)}(\eta_1(z),z_2)}{\eta_2(\eta_1(z),z_2)}\right]\frac{\eta_1^{(1)}(z)}{\eta_1(z)}$$

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

$$H_1(z_2) = G_1(z_1,z_2).$$

Define

$$e_1^{(0)}(y) = y; \quad e_1^{(k)}(y) = e_1\left(e_1^{(k-1)}(y)\right); \quad k \geqslant 1,$$

for $|y| \leqslant 1$.

Iterating (4.11) K times, writing $z_2 = y$,

$$H_1(y) = H_1\left(e_1^{(K+1)}(y)\right)\prod_{k=0}^{K} f_1\left(e_1^{(k)}(y)\right)$$

$$+ H_1(\eta_2(0))\sum_{k=0}^{K} g_1\left(e_1^{(k)}(y)\right)\prod_{l=0}^{k-1} f_1\left(e_1^{(l)}(y)\right)$$

$$+ H_1(0)\sum_{k=0}^{K} h_1\left(e_1^{(k)}(y)\right)\prod_{l=0}^{k-1} f_1\left(e_1^{(l)}(y)\right). \tag{4.12}$$

The next lemma establishes the convergence of (4.12) 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 $|y| \leqslant 1$;

(ii) $\prod_{k=0}^{\infty} f_1\left(e_1^{(k)}(y)\right)$ converges for all y with $|y| \leqslant 1$;

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

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

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 the case of two servers, exponential service times, and exhaustive service,

$$e_1'(1) = \eta_1'(1)\eta_2'(1) = \frac{\rho_1}{2-\rho_1}\frac{\rho_2}{2-\rho_2} = \frac{\rho_1\rho_2}{4-2\rho+\rho_1\rho_2} < 1 \quad \text{iff } \rho < 2.$$

If $e_1'(1) < 1$ then, letting $K \to \infty$ in (4.12),

$$\begin{aligned} H_1(y) = &\prod_{k=0}^{\infty} f_1\left(e_1^{(k)}(y)\right) \\ &+ H_1(\eta_2(0)) \sum_{k=0}^{\infty} g_1\left(e_1^{(k)}(y)\right) \prod_{l=0}^{k-1} f_1\left(e_1^{(l)}(y)\right) \\ &+ H_1(0) \sum_{k=0}^{\infty} h_1\left(e_1^{(k)}(y)\right) \prod_{l=0}^{k-1} f_1\left(e_1^{(l)}(y)\right). \end{aligned} \tag{4.13}$$

Putting $y = 0$ and $y = \eta_2(0)$ in (4.13), 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 the case of an infinite number of servers, deterministic service times, and all the customers present at the start of a visit being 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

$$\begin{aligned} G_1(z_1,z_2) = &\sigma_1(\lambda_2(1-z_2))\eta_2^{(1)}(1,z_2)\eta_1^{(1)}(z) \\ &+ G_1(1,0)\sigma_2(\lambda_2)\sigma_1(\lambda_2(1-z_2))[1-\eta_2^{(1)}(1,z_2)]\eta_1^{(1)}(z) \\ &+ G_1(0,1)\sigma_2(\lambda_1)\sigma_1(\lambda_1+\lambda_2(1-z_2))\eta_2^{(1)}(0,z_2)[1-\eta_2^{(1)}(z)] \\ &+ G_1(0,0)\sigma_2(\lambda_1+\lambda_2)\sigma_1(\lambda_1+\lambda_2(1-z_2))[1-\eta_2^{(1)}(0,z_2)][1-\eta_1^{(1)}(z)]. \end{aligned} \tag{4.14}$$

Putting $z = (1,0)$, $z = (0,1)$, and $z = (0,0)$ in (4.14), 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 $q_1 = 0, q_2 = 1$ described in the previous section, i.e., one of the customers 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.

Equation (4.3) reduces to

$$\begin{aligned} G_1(z_1,z_2) &= G_1(z_1,1)\sigma_2(\gamma(z_1,1))\sigma_1(\gamma(z))\eta_2^{(1)}(z)\frac{\eta_1^{(1)}(z)}{z_1} \\ &\quad + G_1(z_1,0)\sigma_2(\gamma(z_1,0))\sigma_1(\gamma(z))[1-\eta_2^{(1)}(z)]\frac{\eta_1^{(1)}(z)}{z_1} \\ &\quad + G_1(0,1)\sigma_2(\gamma(0,1))\sigma_1(\gamma(0,z_2))\eta_2^{(1)}(0,z_2)\left[1-\frac{\eta_1^{(1)}(z)}{z_1}\right] \\ &\quad + G_1(0,0)\sigma_2(\gamma(0,0))\sigma_1(\gamma(0,z_2))[1-\eta_2^{(1)}(0,z_2)]\left[1-\frac{\eta_1^{(1)}(z)}{z_1}\right]. \end{aligned} \tag{4.15}$$

Setting $z_2 = 0$ and $z_2 = 1$ in (4.15) 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 $2n$ eqs. (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 $E(z^{\mathbf{X}_{ii}}) = F_i(1,\ldots,1,z,1,\ldots,1)$ and $E(z^{\mathbf{Y}_{ii}}) = G_i(1,\ldots,1,z,1,\ldots,1)$, respectively, with z as ith 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 (2.14),

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

with $E(z^{\mathbf{N}_{A,i}}) = (1 - E(z^{\mathbf{X}_{ii}}))/((1-z)E\mathbf{X}_{ii})$. In section 2 we also showed how subsequently the waiting-time distribution may be related to the marginal queue length distribution by using lemma 2.2

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 not, however, 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] M. Ajmone Marsan, S. Donatelli and F. Neri, GSPN models of Markovian multiserver multiqueue systems, Perf. Eval. 11 (1990) 227–240.

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

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

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

[5] B. van Arem, Queueing models for slotted transmission systems, Ph.D. Thesis, Twente University, Enschede (1990).
[6] L.N. Bhuyan, D. Ghosal and Q. Yang, Approximate analysis of single and multiple ring networks, IEEE Trans. Comput. 38 (1989) 1027–1040.
[7] O.J. Boxma, Workloads and waiting times in single-server queues with multiple customer classes, Queueing Systems 5 (1989) 185–214.
[8] O.J. Boxma and W.P. Groenendijk, Two queues with alternating service and switching times, *Queueing Theory and its Applications – Liber Amicorum for J.W. Cohen*, eds. O.J. Boxma and R. Syski (North-Holland, Amsterdam, 1988) pp. 261–282.
[9] S. Browne, E.G. Coffman, jr., E.N. Gilbert and P.E.W. Wright, Gated, exhaustive, parallel service, Prob. Eng. Inf. Sci. 6 (1992) 217–239.
[10] S. Browne and O. Kella, Parallel service with vacations, to appear in Oper. Res.
[11] S. Browne and G. Weiss, Dynamic priority rules when polling with multiple parallel servers, Oper. Res. Lett. 12 (1992) 129–137.
[12] C.H. Chen and L.N. Bhuyan, Design and analysis of multiple token ring networks, *Proc. INFOCOM '88* (1988) pp. 477–486.
[13] S.W. Fuhrmann, Performance analysis of a class of cyclic schedules, Bell Laboratories Technical Memorandum 81-59531-1 (1981).
[14] S.W. Fuhrmann and R.B. Cooper, Stochastic decompositions in the $M/G/1$ queue with generalized vacations, Oper. Res. 33 (1985) 1117–1129.
[15] B. Gamse and G.F. Newell, An analysis of elevator operation in moderate height buildings – II. Multiple elevators, Transp. Res. B16 (1982) 321–335.
[16] A.E. Kamal and V.C. Hamacher, Approximate analaysis of non-exhaustive multiserver polling systems with applications to local area networks, Comput. Netw. ISDN Syst. 17 (1989) 15–27.
[17] E.P.C. Kao and K.S. Narayanan, Analyses of an $M/M/N$ queue with servers' vacations, Eur. J. Oper. Res. 54 (1991) 256–266.
[18] V.V. Karmarkar and J.G. Kuhl, An integrated approach to distributed demand assignment in multiple-bus local networks, IEEE Trans. Comput. 38 (1989) 679–695.
[19] Y. Levy and U. Yechiali, An $M/M/s$ queue with servers' vacations, INFOR 14 (1976) 153–163.
[20] W.M. Loucks, V.C. Hamacher, B.R. Preiss and L. Wong, Short-packet transfer performance in local area ring networks, IEEE Trans. Comput. 34 (1985) 1006–1014.
[21] I.L. Mitrany and B. Avi-Itzhak, A many-server queue with server interruptions, Oper. Res. 16 (1968) 628–638.
[22] R.J.T. Morris and Y.T. Wang, Some results for multi-queue systems with multiple cyclic servers, *Performance of Computer-Communication Systems*, eds. W. Bux and H. Rudin (North-Holland, Amsterdam, 1984) pp. 245–258.
[23] M.F. Neuts and D.M. Lucantoni, A Markovian queue with N servers subject to breakdowns and repairs, Manag. Sci. 25 (1979) 849–861.
[24] T. Raith, Performance analysis of multibus interconnection networks in distributed systems, *Proc. ITC-11*, ed. M. Akiyama (North-Holland, Amsterdam, 1985) pp. 662–668.
[25] J.A.C. Resing, Polling systems and multitype branching processes, Queueing Systems 13 (1993) 409–426.
[26] H. Takagi, A bibliography on the analysis and applications of polling models, *Proc. Int. Workshop on Analysis of Polling Models* (1991).
[27] E.C. Titchmarsh, *The Theory of Functions*, 2nd Ed. (Oxford University Press, London, 1939).
[28] R.W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, 1989).
[29] Q. Yang, D. Ghosal and L.N. Bhuyan, Performance analysis of multiple token ring and multiple slotted ring networks, *Proc. 1986 Comp. Netw. Workshop* (1986) pp. 79–86.

[30] T.I. Yuk and J.C. Palais, Analysis of multichannel token ring networks, *Proc. Int. Conf. Commun. Syst.* (1988) pp. 907–911.
[31] M. Zafirovic-Vukotic, I.G. Niemegeers and D.S. Valk, Performance modelling of slotted ring protocols in HSLAN's, IEEE J. Select. Areas Commun. 6 (1988) 1001–1024.