Annals of Operations Research 48(1994)295–310

Stochastic bounds for a polling system*

O.J. Boxma
CWI (Centre for Mathematics and Computer Science), P.O. Box 4079, 1009 AB Amsterdam, The Netherlands;
Faculty of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

M. Kelbert**
The International Institute for Earthquake Prediction Theory and Mathematical Geophysics, Warshavskoye sh. 79, k2, Moscow 113556, Russia

In this note we consider two queueing systems: a symmetric polling system with gated service at all N queues and with switchover times, and a single-server single-queue model with one arrival stream of ordinary customers and N additional permanently present customers. It is assumed that the combined arrival process at the queues of the polling system coincides with the arrival process of the ordinary customers in the single-queue model, and that the service time and switchover time distributions of the polling model coincide with the service time distributions of the ordinary and permanent customers, respectively, in the single-queue model. A complete equivalence between both models is accomplished by the following queue insertion of arriving customers. In the single-queue model, an arriving ordinary customer occupies with probability p_i a position at the end of the queue section behind the ith permanent customer, $i = 1, \ldots, N$. In the cyclic polling model, an arriving customer with probability p_i joins the end of the ith queue to be visited by the server, measured from its present position.

For the single-queue model we prove that, if two queue insertion distributions $\{p_i, i = 1, \ldots, N\}$ and $\{q_i, i = 1, \ldots, N\}$ are stochastically ordered, then also the workload and queue length distributions in the corresponding two single-queue versions are stochastically ordered. This immediately leads to equivalent stochastic orderings in polling models.

Finally, the single-queue model with Poisson arrivals and $p_1 = 1$ is studied in detail.

Keywords: Polling system, $M/G/1$ queue, permanent customers, stochastic ordering.

1. Introduction

The standard polling system is a single-server multiple-queue system, in

*Part of the research of the first author has been supported by the Esprit BRA project QMIPS.
** Present address: European Business Management School, University of Wales – Swansea, Singleton Park, Swansea SA2 8PP, UK.

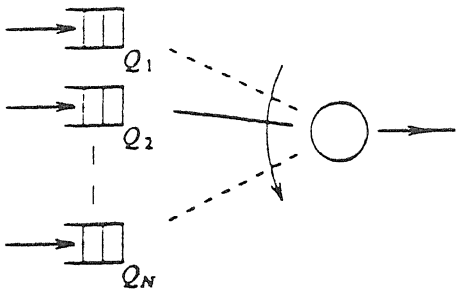

Fig. 1. The polling model.

which the server visits the queues in cyclic order, serving customers at these queues and requiring (possibly zero) switchover times between visits to consecutive queues (cf. fig. 1). Recently, polling systems have received much attention; cf. the extensive survey and list of references in Takagi [10]. The arrival processes at the queues are almost invariably assumed to be independent Poisson processes. Moreover, the position of the server at the epoch of a customer arrival does not influence the number of the queue to which that customer arrives. This is a natural assumption in most applications found in computer communications. However, in case the customers are humans, it seems not unnatural for a customer to take the present position of the server into account and to try and choose a queue that will be visited relatively soon by the server.

In the present note we consider polling systems with some general arrival process of customers to the system (as a whole), and where the server position at the time of arrival of a customer *does* influence the choice of the queue to be joined by the arriving customer. We restrict ourselves mainly to the case of gated service at all queues, viz., at each visit to a queue the server serves exactly those customers found upon its arrival.

Intuitively one expects the polling system to perform relatively well when customers choose queues that are relatively soon to be visited by the server. We show this formally by introducing a stochastic ordering for the queue arrival positions (as measured w.r.t. the server position), and subsequently proving that this leads to a stochastic ordering of workloads (total amount of work to be performed by the server) and of total numbers of customers. In the proof we exploit the relation between (i) polling systems with N queues and switchover times and (ii) single-server *single-queue* systems with externally arriving "ordinary" customers and with N

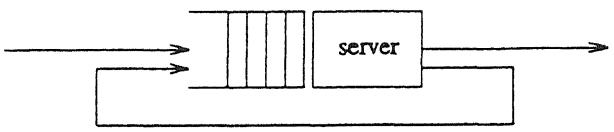

Fig. 2. The $M/G/1$ queue with additional permanent customers.

additional permanently present customers (cf. fig. 2). A permanent customer who has received a service immediately returns to the end of the queue. The relation between these two models has, for a special case, first been observed in [4]. The two models can be made equivalent by equating the total arrival process in both models, by also equating the service time distribution of the customers in the polling model with the service time distribution of the ordinary customers in the single-server single-queue model, and by finally equating the switchover time distribution in the polling model with the service time distribution of the permanent customers. Various rules for the insertion of the ordinary customers between the permanent customers translate into various rules for polling model arrivals that take the server position into account. In particular, the queue insertion probability distribution $P = (p_1, \ldots, p_N)$ indicates that an arriving customer joins the end of the queue section behind the ith permanent customer with probability $p_i, i = 1, \ldots, N$; this translates into a customer in the polling model arriving with probability p_i at the ith queue to be subsequently visited by the server. Stochastic orderings for queue insertion probability distributions in the single-server single-queue model are introduced, leading to the above-mentioned stochastic ordering results for workloads and for numbers of customers (theorem 2.1 and corollary 2.1). The indicated equivalence allows a direct translation to polling models.

One of the advantages of the present study may be that special choices of the arrival rules can lead to relatively simple systems that can be analysed in much detail. This is useful because few detailed polling results are known, even for relatively simple polling systems like those with exhaustive or gated service at all queues and with independent Poisson arrival processes. For example, consider the case that, in the single-server single-queue model with additional permanent customers, ordinary customers join the end of the queue according to a Poisson process, having all permanent customers in front of them. This model has been analysed in [4], where a relatively simple expression for the sojourn time Laplace–Stieltjes Transform (LST) has been derived. From that LST one immediately finds the waiting time LST and hence (PASTA) the workload LST, and finally the LST of the workload of ordinary customers. The results of the present paper show that the latter workload is an upper bound for the workload of the symmetric polling system with gated service to all queues.

Very few stochastic bounds for polling systems are known. Levy et al. [6] compare various service policies w.r.t. total workload, using sample path comparisons. They build a hierarchy of policies, and they show that the exhaustive policy dominates any other policy. Altman et al. [2] show that several performance measures in polling systems are stochastically increasing in arrival rates, service times, and switchover times. Liu et al. [7] try to find dynamic server routing and service policies (taking emptiness, or even exact queue lengths, into account) that stochastically minimize the amount of work and the total number of customers at all times.

The paper is organized as follows. Section 2 contains our main result, a stochastic ordering for workloads in the model with permanent customers. Some

generalizations are also discussed. Subsequently a restriction is made to Poisson arrivals. Three models are described in some detail. Model I is the symmetric cyclic polling model with gated service; model II is the permanent customer model with ordinary (Poisson) customers joining the end of the queue; and model III is the permanent customer model with ordinary (Poisson) customers overtaking all permanent customers except the first one. The mean waiting times of Poisson customers in models I, II and III are presented and compared.

Section 3 is devoted to a detailed analysis of model III. This model can in fact be viewed as an $M/G/1$ queue with gated service and multiple vacations. Our main result for this model is the joint distribution of the numbers of Poisson customers before and after the first permanent customer (before and after the gate), at departure epochs of Poisson customers.

2. A stochastic comparison

Consider two single-server single-queue models with N additional permanently present customers, with as only difference the queue insertion probability distributions $P = (p_1, \ldots, p_N)$ and $Q = (q_1, \ldots, q_N)$ (recall that p_i denotes the probability that an arriving customer joins the end of the queue section behind the ith permanent customer):

$$\sum_{i=1}^{j} p_i \leq \sum_{i=1}^{j} q_i, \quad j = 1, \ldots, N. \tag{2.1}$$

Hence $P \geq_{\text{st}} Q$: P is stochastically larger than Q, cf. Ross [8]. The arrival process of ordinary customers is an arbitrary stochastic process, which is the same for both models. Similarly, the service times of ordinary customers are independent generally distributed stochastic variables, with identical distributions in both models. A similar statement holds for the service times of permanent customers. The interarrival, service and switchover processes are assumed to be independent. It is assumed that the traffic characteristics of ordinary customers are such that the limiting distributions of workloads and of total numbers of ordinary customers in both models exist and are equal to the stationary distributions.

Denote by $\boldsymbol{V}^{(P)}$ ($\boldsymbol{V}^{(Q)}$) the steady-state total workload of ordinary customers in model P (Q), the model with queue insertion probability distribution P (Q). Our main result is:

THEOREM 2.1

$$\boldsymbol{V}^{(P)} \geq_{\text{st}} \boldsymbol{V}^{(Q)}. \tag{2.2}$$

Proof

The proof of this theorem is based on classical coupling arguments. In fact, we shall compare sample paths for models P and Q, and we shall show that with an appropriate coupling the workload of ordinary customers in model P majorizes the workload of ordinary customers in model Q for each sample path. Let us consider the evolution of both models starting with no ordinary customers at $t = 0$. We shall prove that at any moment $t \geq 0$,

$$V^{(P)}(t) \geq_{\text{st}} V^{(Q)}(t),$$

with $V^{(P)}(t)$ ($V^{(Q)}(t)$) the total workload of ordinary customers at time t in model P (Q) (note that this is a stronger result than (2.2)). As the ergodicity condition is fulfilled, we obviously have the same inequality

$$V^{(P)} \geq_{\text{st}} V^{(Q)}$$

for stationary versions of those workloads.

The stochastic ordering of P and Q implies a stochastic ordering of the stochastic vectors of queue insertion positions $\bar{r}^{(P)} = (r_1^{(P)}, r_2^{(P)}, \cdots)$ and $\bar{r}^{(Q)} = (r_1^{(Q)}, r_2^{(Q)}, \cdots)$. According to Strassen's lemma, there exist stochastic vectors $\bar{r}^{(P)}$ and $\bar{r}^{(Q)}$ such that $Pr\{\bar{r}^{(P)} \geq \bar{r}^{(Q)}\} = 1$. Consider a realization ω of the input sequence of arrival times, service times of ordinary and permanent customers and queue insertion positions $(r_1^{(P)}, r_2^{(P)}, \ldots)$ for model P. Take the following coupling: arrival times of successive ordinary customers in both models are identical; the service time of the jth service of an ordinary customer (respectively of a permanent customer) in both models is identical, $j = 1, 2, \ldots$; and $r_j^{(P)} = r_j^{(Q)}$ for all $j \neq h$, while $r_h^{(P)} > r_h^{(Q)}$. Then obviously $V^{(P)}(t)$ equals $V^{(Q)}(t)$ with the exception of a period starting with the arrival of the hth ordinary customer, during which period $V^{(P)}(t) > V^{(Q)}(t)$.

Subsequently allow $r_k^{(P)} > r_k^{(Q)}$ for a second index $k \neq h$, etc.; iterating this procedure we end up with $V^{(P)}(t) \geq V^{(Q)}(t)$ for all $t \geq 0$. The proof is concluded by removing the conditioning on ω. □

Remark 2.1

The chosen coupling of the service times in the proof of theorem 2.1 immediately shows that the numbers of ordinary customers in models P and Q are also stochastically ordered. Such an ordering does not hold for *sojourn times* of ordinary customers. Counterexamples can be easily constructed, exploiting the fact that ordinary customers are not necessarily served in order of arrival.

The above stochastic ordering results can be adapted and generalized in various ways. A brief discussion of three such possibilities is given below.

(i) In theorem 2.1 it is assumed that an arriving customer occupies, with probability p_i, a position *at the end* of the queue section behind the ith permanent customer. The proof of the theorem shows that the exact position occupied within the ith queue section is not relevant for its result.

(ii) In theorem 2.1 an arriving customer can only occupy a position in the queue sections behind the ith permanent customer, $i = 1, \ldots, N$; in the framework of polling, this corresponds to gated service. Let us now also allow arriving customers to occupy a position in the queue section *before* the first permanent customer (if that permanent customer is not in service). For example, if an arriving customer joins the queue section before (respectively after) the ith permanent customer with probability $1/N, i = 1, \ldots, N$, when upon its arrival an ordinary (respectively permanent) customer is in service, then the corresponding polling model is the symmetric cyclic polling model with *exhaustive* service. If *each* arriving customer joins the queue section before the first permanent customer, then a single-server queue with exhaustive service and multiple vacations results. Generally, we could have queue section insertion probabilities $p_i^{(o)}, i = 0, \ldots, N$, when an arriving customer finds an ordinary customer in service, and queue section insertion probabilities $p_i^{(p)}, i = 1, \ldots, N$, when an arriving customer finds a permanent customer in service. If the corresponding $P^{(o)}$ is stochastically larger than both $Q^{(o)}$ and $Q^{(p)}$, and similarly $P^{(p)}$ is stochastically larger than both $Q^{(o)}$ and $Q^{(p)}$, then again stochastic orderings of workloads and numbers of customers can be proved. For example, taking $p_i^{(p)} = q_i^{(p)} = 1/N, i = 1, \ldots, N, p_i^{(o)} = q_{i-1}^{(o)} = 1/N, i = 1, \ldots, N$, leads to the result that the workload in the symmetric polling system with exhaustive service is stochastically smaller than the workload in the symmetric polling system with gated service (a result that has already been obtained in [6]).

In principle one can go even further, and allow the possibility that an arriving customer has to wait several cycles before receiving service. For this purpose one has to introduce more permanent customers than there are queues in the corresponding polling model.

(iii) In the framework of polling it might be interesting to allow a more general influence of the server position on the choice of queue for an arriving customer, by assigning newly arriving customers with probability p_{ik} to $Q_{(i+k) \bmod N}$ when the server is at Q_i. In the single-server single-queue model this corresponds to making a distinction between the various permanent customers. Polling models where the position of the server influences the choice of the queue at which a new customer arrives have hardly been considered so far. At CWI a detailed study of such models is being started, including the existence of conservation laws.

In the remainder of this paper we restrict ourselves almost exclusively to three special cases of the above model, *with ordinary customers arriving according to a*

Poisson process. These models will be called model I, II, and III, and are described below in some detail.

2.1. DESCRIPTION OF MODEL I (CF. FIG. 1)

Model I is a cyclic polling system with N queues, $Q_1, \ldots, Q_N$, where each queue has an infinite buffer capacity to store waiting customers. Customers arrive at all queues according to independent Poisson processes. The arrival intensity at Q_i is $\lambda/N, i = 1, \ldots, N$. The service times of the customers at all queues are i.i.d. stochastic variables with general distribution $B(\cdot)$, first moment β and second moment $\beta^{(2)}$ and Laplace–Stieltjes transform $\beta\{\cdot\}$. The total offered traffic to the sytem is $\rho = \lambda\beta$. The queues are attended by a single server who visits the queues in a fixed cyclic order. The switchover times of the server between any two consecutive queues Q_i, Q_{i+1} are i.i.d. stochastic variables with general distribution $S(\cdot)$, first moment s and second moment $s^{(2)}$ and Laplace–Stieltjes transform $\sigma\{\cdot\}$. The interarrival, service and switchover processes are assumed to be independent. The server serves each queue according to the gated discipline, and serves customers within each queue in FIFO order. The server keeps switching in an empty system. It is well-known that in this model $\rho < 1$ is a necessary and sufficient ergodicity condition (Takagi [9]). As argued in this paper, model I can also be viewed as a single-server single-queue model with N additional permanently present customers, the latter ones having service time distribution $S(\cdot)$, and with the ordinary (Poisson) customers occupying the position at the end of the queue section behind the ith permanent customer with probability $p_i = 1/N, i = 1, \ldots, N$.

2.2. DESCRIPTION OF MODEL II (CF. FIG. 2)

Model II differs from the permanent customer version of model I in only one respect. The queue section insertion probability distribution is $(0, 0, \ldots, 1)$: The Poisson customers join the very end of the queue, behind all permanent customers. Figure 3 indicates the queue composition of this model.

It was first observed in Boxma and Cohen [4] that model II can be viewed as a – rather special – cyclic polling model with N queues and gated service at all queues. The service times of the permanent customers correspond to the switchover times of the server between successive queues. To take into account that in the $M/G/1$ model

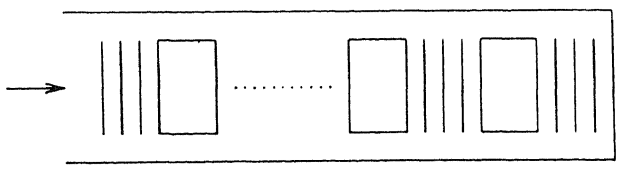

Fig. 3. Queue composition in model II (a box denotes a permanent customer).

there is really only one queue, one has to assume that arrivals at a particular queue of the special polling model are only possible during the server visit to that queue and during the subsequent switchover time; thereafter, arrivals can only take place at the next queue, etc. This "discriminatory" arrival process at the queues of the special polling model, with customers arriving at what is apparently the worst possible queue, makes it intuitively clear that the workload of Poisson customers in model II is stochastically larger than the total workload in model I.

The permanent customer version of model II has been analysed in detail in [4]. It has been shown that $\rho < 1$ is a necessary and sufficient ergodicity condition. Among other things, an explicit expression is obtained for the generating function of the queue length distribution of the Poisson customers at departure epochs of Poisson customers. This immediately yields an expression for the LST of the sojourn and waiting time distributions of the Poisson customers, from which moments can be easily obtained.

2.3. DESCRIPTION OF MODEL III

Model III deviates from model II in only one respect. The queue section insertion probability distribution is $(1, 0, 0, \ldots, 0)$: The Poisson customers join the queue in the order of their arrival, but overtaking all permanent customers except for the first one. Figure 4 indicates the queue composition of this model. Just like model II, model III is equivalent to a cyclic polling model with N queues and gated service at all queues, in which the arrival process is special: arrivals at a particular queue only take place during the server visit to the *previous* queue and during the subsequent switchover time. Customers hence arrive at what is from their point of view the *best* possible queue, making it intuitively clear that the workload of Poisson customers in model III is stochastically smaller than the total workload in model I.

Let V_I, V_{II} and V_{III} denote the steady-state workloads of Poisson customers in models I, II and III. The customer insertion probabilities for models I, II and III (respectively: $p_i = 1/N, i = 1, \ldots, N$ for model I, $p_N = 1$ for model II, and $p_1 = 1$ for model III) immediately lead to the following corollary of theorem 2.1 (obviously the corollary holds for a general arrival process of the ordinary customers).

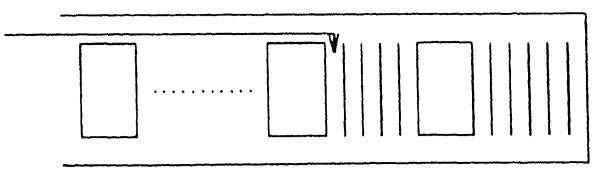

Fig. 4. Queue composition in model III (a box denotes a permanent customer).

COROLLARY 2.1

$$V_{II} \geq_{\text{st}} V_I \geq_{\text{st}} V_{III}. \tag{2.3}$$

Remark 2.2

The fact that $V_{II} \geq_{\text{st}} V_{III}$, combined with the PASTA property and the FIFO order of service for Poisson customers in models II and III, implies a stochastic ordering of waiting times in models II and III. However, such a stochastic ordering does not hold w.r.t. model I. It is easy, though, to rank the *mean* waiting times in models I, II and III. In model I (cf. Takagi [9]),

$$EW_I = \frac{\lambda\beta^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}(N-1+2\rho). \tag{2.4}$$

In model II, one can obtain the mean of Z_{II}, the number of Poisson customers left behind after the departure of a Poisson customer, from (2.8) of [4]; using the obvious relation $EZ_{II} = \lambda(EW_{II} + \beta)$ we find:

$$EW_{II} = \frac{\lambda\beta^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{(1-\rho)}(N-1+\rho). \tag{2.5}$$

Some reflection shows that Poisson customers do not see a difference between model III and the $N = 1$ variant of model II. The latter model is an $M/G/1$ queue with one permanent customer, and can also be viewed as an $M/G/1$ queue with gated service and multiple vacations. The mean waiting time of Poisson customers now follows either from (2.5) with $N = 1$ or from section 2.5 of [11]:

$$EW_{III} = \frac{\lambda\beta^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{\rho s}{1-\rho}. \tag{2.6}$$

Hence

$$EW_I - EW_{III} = EW_{II} - EW_I = \frac{s}{2(1-\rho)}(N-1) \geq 0, \tag{2.7}$$

with equality when $s = 0$ or $N = 1$. Application of Little's formula immediately gives similar results for mean numbers of customers.

Remark 2.3

Consider the following variant of model I, the cyclic polling model: model RP, a random polling model in which the next queue to be visited by the server is

Q_i with probability $1/N$, $i = 1, 2, \ldots, N$, regardless of which queues have previously been visited. Assume that all traffic characteristics and switchover time distributions are the same as in model I. Comparison of (2.5) and formula (5.37) of [5] reveals the intriguing fact that *the mean waiting time* EW_{RP} *at all queues in model RP equals* EW_{II}. In model II, the server visits all queues exactly once in every set of N consecutive visits, according to a cyclic pattern, and the customers arrive "in the worst possible queue". In model RP the customer arrival process uses no information about the server position, and the server visits all queues *on the average* one out of N times, but according to a completely random pattern.

The equality $EW_{RP} = EW_{II}$ can easily be verified by using the theory of work decomposition for single server queueing systems with service interruptions (vacations, switchover times; cf. [3]). Denote by V_{RP} the steady-state amount of work in model RP, and denote by Y_{RP} (Y_{II}) the steady-state amount of work in model RP (workload of Poisson customers in model II) at a switchover epoch of the server. Finally, denote by $V_{M/G/1}$ the steady-state amount of work in the corresponding $M/G/1$ queue *without* switchover times (an $M/G/1$ queue with arrival rate λ and service time distribution $B(\cdot)$). It follows from [3] that

$$EV_{RP} = EV_{M/G/1} + EY_{RP},$$

$$EV_{II} = EV_{M/G/1} + EY_{II}.$$

Next we show that $EY_{RP} = EY_{II}$ and hence $EV_{RP} = EV_{II}$; this rapidly implies the equality of the mean waiting times in both models: $EW_{RP} = EW_{II}$, cf. formula (3.5) of [3]. Note that both EY_{RP} and EY_{II} can be written as the sum of $\rho s^{(s)}/2s$ (the mean amount of work that has arrived in the past part of the switchover interval under consideration) and the mean amount of work present in the system at the *beginning* of that switchover interval. In model II, all this work is gathered in the past cycle, minus one switchover period; its mean equals (with EC the mean cycle time) $\rho[EC - s] = \rho s(N - 1 + \rho)/(1 - \rho)$. Similarly, in model RP each queue was last left behind by the server on the average N visits ago, hence $Ns/(1 - \rho) - s = s(N - 1 + \rho)/(1 - \rho)$ time units ago, and has since then acquired ρ/N work per unit of time. Hence the mean amounts of work in both models at the beginning of a switchover interval are the same, and so are EY_{RP} and EY_{II}.

3. Analysis of model III

In this section we analyse the queue length process in model III, viz., the $M/G/1$ queue with additional permanent customers, in which arriving Poisson customers join the queue in order of arrival and occupy the position immediately ahead of the second permanent customer. As observed in remark 2.2, this model is equivalent to an $M/G/1$ queue with only one permanent customer (model II with $N = 1$), and hence also to an $M/G/1$ queue with multiple vacations and gated service. The latter model has been extensively studied. Our main contribution is an

exact analysis of the *joint distribution* of the numbers of customers ahead of and behind the (first) permanent customer, at departure epochs of Poisson customers. The special structure of the model allows us to solve the two-dimensional functional equation for the generating function of the joint queue length distribution. The functional equation is of a type that is not completely uncommon in branching-type queueing models. The queue length analysis easily leads to the marginal and total queue length distributions, and to the LST and mean of the waiting times and workload. The distribution of the total queue length can also be found in Takagi [11, section 2.5] and in [4]. The latter paper considers also the joint queue length distribution at service completion epochs of *arbitrary* customers, Poisson and permanent alike. A detailed study of the $M/G/1$ queue with multiple vacations and gated service is given in Takine and Hasegawa [12]. They present the time-dependent analysis of the numbers of customers before and after the gate at time t. It should be noted that the joint limiting distribution of these numbers differs from the limiting distribution at departure epochs of Poisson customers. Another related paper is Ali and Neuts [1]. In their model customers arrive according to a Poisson process, and wait for service in a two-stage queue. The first stage is a waiting room; the second stage resides in the service room. Whenever the service room becomes empty, it is replenished by the transfer of all customers in the waiting room and the addition of a positive random number of overhead customers. These overhead customers play the role of our permanent customers, while the presence of the waiting room corresponds to the insertion of Poisson customers *after* the first permanent customer. In the model of Ali and Neuts, overhead customers have the same service time distribution as Poisson customers. Under this assumption, they determine, a.o., the joint distribution of the numbers of customers in both stages.

We now present our analysis of model III. Consider the epoch of the nth departure of a Poisson customer. Denote by $Z_n^{(1)}$ and $Z_n^{(2)}$ the numbers of Poisson customers at this epoch before and after the first permanent customer; the last $N-1$ permanent customers always form the tail of the queue. We have

if $Z_n^{(1)} > 0$: (3.1)

$$Z_{n+1}^{(1)} = Z_n^{(1)} - 1,$$
$$Z_{n+1}^{(2)} = Z_n^{(2)} + \nu_{n+1};$$

if $Z_n^{(1)} = 0$:

$$Z_{n+1}^{(1)} = Z_n^{(2)} + \mu_{n+1} - 1 \quad \text{if } Z_n^{(2)} > 0,$$
$$Z_{n+1}^{(1)} = \hat{\mu}_n^{(1)} - 1 \quad \text{if } Z_n^{(2)} = 0,$$
$$Z_{n+1}^{(2)} = \nu_{n+1}.$$

Here μ_n (ν_n) denotes a s.v. with distribution that of the number of Poisson arrivals during the service of a permanent (Poisson) customer; $\hat{\mu}_n$ denotes the same quantity as μ_n, but under the condition that the number of arrivals is *positive*. Note that

$$\begin{aligned} &E[r^{\nu_n}] = \beta\{\lambda(1-r)\}, \quad E[r^{\mu_n}] = \sigma\{\lambda(1-r)\}, \\ &E[r^{\hat{\mu}_n}] = \hat{\sigma}\{\lambda(1-r)\} := [\sigma\{\lambda(1-r)\} - \sigma\{\lambda\}]/[1-\sigma\{\lambda\}]. \end{aligned} \tag{3.2}$$

Denote by $F_n(r_1, r_2)$ the generating function of the joint distribution of $\boldsymbol{Z}_n^{(1)}$ and $\boldsymbol{Z}_n^{(2)}$. It can be shown that $\lim_{n\to\infty} F_n(r_1, r_2) = F(r_1, r_2)$ exists when $\rho < 1$, with $F(r_1, r_2)$ denoting the generating function of the joint stationary distribution of $\boldsymbol{Z}_n^{(1)}$ and $\boldsymbol{Z}_n^{(2)}$. From the set of recurrence relations (3.1) it follows by standard analysis that $F(r_1, r_2)$ satisfies the following functional equation for $|r_1| \le 1, |r_2| \le 1$:

$$\begin{aligned} F(r_1, r_2) = &\frac{1}{r_1}\beta\{\lambda(1-r_2)\}[F(r_1, r_2) - F(0, r_2)] \\ &+ \frac{1}{r_1}\sigma\{\lambda(1-r_1)\}\beta\{\lambda(1-r_2)\}[F(0, r_1) - F(0,0)] \\ &+ \frac{1}{r_1}\hat{\sigma}\{\lambda(1-r_1)\}\beta\{\lambda(1-r_2)\}F(0,0). \end{aligned} \tag{3.3}$$

Hence, for $|r_1| \le 1, |r_2| \le 1$,

$$\begin{aligned} F(r_1, r_2) = &[r_1 - \beta\{\lambda(1-r_2)\}]^{-1} \\ &\times [-\beta\{\lambda(1-r_2)\}F(0, r_2) + \sigma\{\lambda(1-r_1)\}\beta\{\lambda(1-r_2)\}F(0, r_1) \\ &+ [\hat{\sigma}\{\lambda(1-r_1)\} - \sigma\{\lambda(1-r_1)\}]\beta\{\lambda(1-r_2)\}F(0,0)]. \end{aligned} \tag{3.4}$$

Note that for every r_2 with $|r_2| \le 1$, the denominator of the right-hand side of (3.4) has exactly one zero $r_1 = \beta\{\lambda(1-r_2)\}$, and $|r_1| \le 1$. Since $F(r_1, r_2)$ is an analytic function in r_1 and r_2 for $|r_1| \le 1$, $|r_2| \le 1$, the numerator of the right-hand side of (3.4) must be zero for all these zeros $r_1 = \beta\{\lambda(1-r_2)\}$. Hence, defining

$$\delta(r) := \beta\{\lambda(1-r)\}, \quad |r| \le 1,$$

it follows, using (3.2), that for $|r| \le 1$,

$$\begin{aligned} F(0, r) &= \sigma\{\lambda(1-\delta(r))\}F(0, \delta(r)) + [\hat{\sigma}\{\lambda(1-\delta(r))\} - [\sigma\{\lambda(1-\delta(r))\}]F(0,0) \\ &= \sigma\{\lambda(1-\delta(r))\}F(0, \delta(r)) + [\sigma\{\lambda(1-\delta(r))\} - 1]\frac{\sigma\{\lambda\}}{1-\sigma\{\lambda\}}F(0,0). \end{aligned} \tag{3.5}$$

We shall determine $F(0,r)$ – and hence finally $F(r_1,r_2)$ from (3.3) – by iteratively solving (3.5), successively replacing r by $\delta(r), \delta(\delta(r)), \cdots$ in the left-hand side. Introduce for $|r| \leq 1$,

$$\delta^{(0)}(r) := r,$$
$$\delta^{(k)}(r) := \delta(\delta^{(k-1)}(r)), \qquad k = 1,2,\ldots,$$
$$\phi^{(k)}(r) := \lambda(1-\delta^{(k)}(r)), \qquad k = 0,1,\ldots.$$

The kind of functional equation (3.5) is not uncommon in branching-type queueing models like the one under consideration. Note that $\delta^{(k)}(r)$ can be viewed as the generating function of the number of kth generation offspring of a single element in the 0th generation, with $\beta\{\lambda(1-r)\}$ the generating function of the branching distribution of a single element. Similar to the analysis on p. 180 of [4] one can show that successive iteration of (3.5) converges iff $\rho < 1$. For $\rho < 1$ one obtains:

$$F(0,r) = F(0,1)\prod_{j=1}^{\infty}\sigma\{\phi^{(j)}(r)\} + F(0,0)\frac{\sigma\{\lambda\}}{1-\sigma\{\lambda\}}\sum_{j=1}^{\infty}\Big[\sigma\{\phi^{(j)}(r)\}-1\Big]\prod_{h=1}^{j-1}\sigma\{\phi^{(h)}(r)\}$$
$$= F(0,1)\prod_{j=1}^{\infty}\sigma\{\phi^{(j)}(r)\} + F(0,0)\frac{\sigma\{\lambda\}}{1-\sigma\{\lambda\}}\left[\prod_{h=1}^{\infty}\sigma\{\phi^{(h)}(r)\}-1\right]. \tag{3.6}$$

It remains to determine $F(0,1)$ and $F(0,0)$. Substitution of $r=0$ into (3.6) yields one linear relation between $F(0,0)$ and $F(0,1)$. A second linear relation between these quantities is obtained by substituting $r_1 = r_2 = r$ into (3.4):

$$F(r,r) = (1-\rho)\frac{(1-r)\beta\{\lambda(1-r)\}}{\beta\{\lambda(1-r)\}-r}$$
$$\times\left[\frac{1-\sigma\{\lambda(1-r)\}}{1-r}\frac{F(0,r)}{1-\rho} + \frac{\sigma\{\lambda(1-r)\}-\hat{\sigma}\{\lambda(1-r)\}}{1-r}\frac{F(0,0)}{1-\rho}\right], \quad |r| \leq 1. \tag{3.7}$$

Taking $r=1$ in (3.7) yields the required second linear relation between $F(0,1)$ and $F(0,0)$:

$$1 = \left[\lambda s F(0,1) + \lambda s\frac{\sigma\{\lambda\}}{1-\sigma\{\lambda\}}F(0,0)\right]/[1-\rho]. \tag{3.8}$$

This formula can easily be interpreted by rewriting it into

$$\frac{1-\rho}{\lambda s} = [F(0,1)-F(0,0)] + F(0,0)/[1-\sigma\{\lambda\}],$$

and observing that both the left-hand side and the right-hand side represent the ratio of services given to permanent and Poisson customers. Indeed, $(1-\rho)/s$ permanent customers are served per unit of time, as oppposed to λ Poisson customers; and $1/(1-\sigma\{\lambda\})$ denotes, starting from a system with no Poisson customers present, the average number of permanent services before a new Poisson arrival must be served.

Using (3.6) and (3.8) we can rewrite (3.7):

$$F(r,r) = (1-\rho)\frac{(1-r)\beta\{\lambda(1-r)\}}{\beta\{\lambda(1-r)\}-r}\left[\frac{1-\sigma\{\lambda(1-r)\}}{(1-r)\lambda s}\prod_{j=1}^{\infty}\sigma\{\phi^{(j)}(r)\}\right], \quad |r| \le 1. \tag{3.9}$$

The decomposition structure of (3.9) should be noted; the left-hand side of (3.9) denotes the generating function of the distribution of the total number of Poisson customers in the system *with* permanent customers, and the term outside the square brackets in the right-hand side of (3.9) represents the generating function of the distribution of the number of customers in the corresponding $M/G/1$ queue *without* permanent customers. Note that $F(r,r) = E[\exp(-\lambda(1-r)W_{III})]\beta\{\lambda(1-r)\}$, which determines the LST of the waiting time distribution of Poisson customers. The PASTA property and the FIFO order of service of the Poisson customers imply that the LST of the waiting time distribution equals the LST of the total workload distribution. We refer the reader to section 2.5 of Takagi [11] for further discussions of the queue length and waiting time processes in this model (viewed as an $M/G/1$ queue with multiple vacations and gated service).

In principle one can calculate the correlation between the two queue lengths before and after the permanent customer. We refrain from presenting the results of this lengthy calculation; instead, we turn to the marginal limiting distributions of $Z_n^{(1)}$ and $Z_n^{(2)}$. From (3.4), (3.6) and (3.8) we obtain their generating functions, for $|r| \le 1$:

$$\begin{aligned} F(r,1) &= \frac{1-\rho}{\lambda s}\frac{\prod_{j=0}^{\infty}\sigma(\phi^{(j)}(r))-1}{r-1}, \\ F(1,r) &= \frac{\beta\{\lambda(1-r)\}(1-r)}{1-\beta\{\lambda(1-r)\}}\frac{F(0,1)-F(0,r)}{1-r}. \end{aligned} \tag{3.10}$$

Using (3.6) and (3.8), the latter relation can be rewritten into

$$F(1,r) = \frac{\beta\{\lambda(1-r)\}}{1-\beta\{\lambda(1-r)\}}\frac{1-\rho}{\lambda s}\frac{\prod_{j=1}^{\infty}\sigma(\phi^{(j)}(r))-1}{r-1}. \tag{3.11}$$

In [4] it has been observed that $\prod_{j=0}^{\infty} \sigma(\phi^{(j)}(r))$ is the GF of the queue length distribution at the end of the service of a (the) permanent customer; apparently the right-hand side of (3.10) is the GF of the corresponding overshoot distribution. The interpretation is clear: at the end of the service of a permanent customer, the block of Poisson customers after this permanent customer is "complete", and becomes the block of Poisson customers *ahead* of the (first) permanent customer. The size of this block subsequently reduces after each service completion. At a departure epoch of a Poisson customer, $F(r, 1)$ gives the GF of the remainder of such a block.

It is easy to determine EZ_{III}, the mean number of Poisson customers left behind after the departure of a Poisson customer. Note that EZ_{III} equals the derivative of $F(r, r)$ at $r = 1$. Differentiating both sides of (3.9) we find

$$EZ_{III} = \rho + \frac{\lambda^2 \beta^{(2)}}{2(1-\rho)} + \frac{\lambda s^{(2)}}{2s} + \frac{\lambda \rho s}{1-\rho}. \tag{3.12}$$

The mean waiting time EW_{III} of Poisson customers in model III follows from the obvious relation $EZ_{III} = \lambda(EW_{III} + \beta)$; we find, cf. (2.6):

$$EW_{III} = \frac{\lambda \beta^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{\rho s}{1-\rho}. \tag{3.13}$$

Remark 3.1

Using the work decomposition reasoning at the end of remark 2.3, one can readily generalize (3.13) and (2.5) to the case of a customer insertion that occurs with probability one at the end of the queue section behind the ith permanent customer: The mean waiting time $EW^{(i)}$ for the case of $p_i = 1$ becomes:

$$EW^{(i)} = \frac{\lambda \beta^{(2)}}{2(1-\rho)} + \frac{s^{(2)}}{2s} + \frac{s}{1-\rho}(i - 1 + \rho). \tag{3.14}$$

Note that a model with N permanent customers with $p_i = 1$ corresponds (as far as the Poisson customers is concerned) to a model with only i permanent customers, with insertion of arriving Poisson customers at the end of the line. This observation allows us to deduce from [4] that the generating function of the distribution of the total number of Poisson customers $X_{III}^{(i)}$ in the case $p_i = 1$ is given by (cf. (3.9)):

$$E\left[r^{X_{III}^{(i)}}\right] = (1-\rho)\frac{(1-r)\beta\{\lambda(1-r)\}}{\beta\{\lambda(1-r)\} - r} \times \left[\frac{1-\sigma\{\lambda(1-r)\}}{(1-r)\lambda s}\left[\prod_{j=1}^{\infty} \sigma\{\phi^{(j)}(r)\}\right]^i\right] / \sigma\{\lambda(1-r)\}, \quad |r| \leq 1. \tag{3.15}$$

The generating function of the total queue length distribution in a symmetric polling model with gated service should be accurately represented by a weighted sum of the expression in (3.15) for $i = 1, \ldots, N$, with weight factors $1/N$.

Acknowledgements

The authors are grateful to H. Levy for an observation that has led to remark 2.3, to G.M. Koole for valuable comments about stochastic orderings, and to S.C. Borst and M.B. Combé for interesting discussions.

References

[1] O.M.E. Ali and M.F. Neuts, A service system with two stages of waiting and feedback of customers, J. Appl. Prob. 21 (1984) 404–413.
[2] E. Altman, P. Konstantopoulos and Z. Liu, Stability, monotonicity and invariant quantities in general polling systems, Queueing Syst. 11 (1992) 35–57.
[3] O.J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, Queueing Syst. 5 (1989) 185–214.
[4] O.J. Boxma and J.W. Cohen, The $M/G/1$ queue with permanent customers, IEEE J. Sel. Areas Commun. SAC-9 (1991) 179–184.
[5] O.J. Boxma and J.A. Weststrate, Waiting times in polling systems with Markovian server routing, in: *Messung, Modellierung und Bewertung von Rechensystemen und Netzen*, eds. G. Stiege and J.S. Lie (Springer, Berlin, 1989) pp. 89–104.
[6] H. Levy, M. Sidi and O.J. Boxma, Dominance relations in polling systems, Queueing Syst. 6 (1990) 155–171.
[7] Z. Liu, Ph. Nain and D.F. Towsley, On optimal polling policies, Queueing Syst. 11 (1992) 59–83.
[8] S.M. Ross, *Stochastic Processes* (Wiley, New York, 1983).
[9] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).
[10] H. Takagi, Queueing analysis of polling models: an update, in: *Stochastic Analysis of Computer and Communication Networks*, ed. H. Takagi (North-Holland, Amsterdam, 1990) pp. 267–318.
[11] H. Takagi, *Queueing Analysis. Vol. 1: Vacation and Priority Systems, Part 1* (North-Holland, Amsterdam, 1991).
[12] T. Takine and T. Hasegawa, On the $M/G/1$ queue with multiple vacations and gated service discipline, J. Oper. Res. Soc. Japan (1992), to appear.