A polling system with a dormant server

S.C. Borst

Department of Operations Research, Statistics, and System Theory

Report BS-R9313 June 1993

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

P.O. Box 4079, 1009 AB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

A Polling System with a Dormant Server

S.C. Borst
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Abstract

We study a polling system with a dormant server, i.e., a polling system in which the server may be allowed to make a halt at a queue when there are no customers present in the system. We derive a pseudo-conservation law for a general model and use it to compare the dormant and the non-dormant server case. We further address the question at which queues the server should make a halt to minimize the mean total amount of work in the system. Subsequently we direct the attention in particular to a globally gated polling system in which the server makes a halt at its home base when there are no customers present in the system. For that system we derive an explicit expression for the LST of the cycle time distribution, for the LST of the waiting time distribution at each of the queues, and for the pgf of the joint queue length distribution at polling epochs. The waiting time at each of the queues is shown to be smaller (in the increasing convex ordering sense) than in the non-dormant server case.

1991 Mathematics Subject Classification: 60K25, 68M20.
Keywords & Phrases: dormant server, globally gated service, polling system, pseudo-conservation law, queue lengths, waiting times.

1 INTRODUCTION

A polling system basically consists of several queues attended to by a single server. The service discipline prescribes which customers are to be served during a visit to a queue, in other words, it dictates the server when to move from one queue to another. The server routing dictates the server from which queue to which queue to move. Moving from one queue to another typically requires a non-zero switch-over time. Polling systems arise quite naturally in modelling situations in which several users compete for service from a single common server. Thus polling systems have found a variety of applications in the areas of computer networks, telecommunication, manufacturing, and maintenance, cf. Takagi [18], [20] and Levy & Sidi [15]. Motivated by the variety of applications, numerous studies have been devoted to the analysis of polling systems, cf. [18], [20].

In the present paper we study a polling system with a *dormant* server, i.e., a polling system in which the server may be allowed to make a halt at a queue when there are no customers present in the system. Usually in the polling literature the server is assumed never to idle, in other words, to be switching when not working. In particular the server is assumed to be switching when there are no customers present in the system. As a rare exception, Eisenberg [11] analyzes a two-queue model with either alternating priority (the exhaustive service

Report BS-R9313
ISSN 0924-0659
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

discipline at both queues) or strict priority, in which the server remains idling at a queue when there are no customers present in the system. Eisenberg [12] analyzes a model with an arbitrary number of queues and the exhaustive service discipline at all queues, in which the server does *not* idle. However, as indicated by Eisenberg [private communication], an adapted version of the solution method in [12] may be used to deal with a model in which the server makes a halt at some of the queues when the system is empty. The analysis in [12] starts from an equation which expresses that each time a visit beginning or a service completion occurs, simultaneously also either a service beginning or a visit completion occurs. To deal with a model in which the server makes a halt at some of the queues when the system is empty, this equation should be modified, as a service completion that leaves the entire system empty then does not coincide with a visit completion. The modification complicates the analysis of the model, but the outline of the solution method in [12] may still be used. In particular the analysis yields the probabilities that the server is idling at the various queues. Similar remarks hold for the gated service discipline. Liu, Nain, & Towsley [16] identify polling policies, allowing idling as a possible action, that stochastically minimize the total amount of work in the system at an arbitrary epoch. They show that optimal policies are exhaustive, greedy, and in a symmetrical system in addition patient, i.e., the server should neither switch nor idle when at a non-empty queue and in a symmetrical system the server should remain idling at a queue when the entire system is empty. Blanc & Van der Mei [3] use the power-series algorithm to analyze the performance of a system in which the server may be allowed to make a halt at a queue when the entire system is empty. They find that the performance may improve considerably by allowing the server to make a halt at a queue, especially in a light traffic situation.

One reason why usually in the polling literature the server is nevertheless assumed never to idle, may be that the option of idling in general slightly complicates the operation of the system under consideration. If at all technically feasible, some mechanism is needed to control the server and to keep track of the customers present in the system. Consequently, the option of idling in general also slightly complicates the analysis of the system under consideration. Another reason may be that the option of idling will have the biggest impact in a light traffic situation - hence when the system performance will be satisfactory anyhow.

However, quite often there are very sound reasons for letting the server stop switching when there are no customer present in the system. In many situations some mechanism to control the server and to keep track of the customers present in the system is needed anyhow. The option of idling then enters the picture quite naturally. In manufacturing and maintenance environments e.g. one usually requires already some kind of supporting system to schedule the jobs. In such situations it makes sense to let the server make a halt at a queue when the entire system is empty, rather than to let the server needlessly circle around. One option is then allowing the server to make a halt at all of the queues, i.e., to stop switching as soon as the entire system is empty. Another option is allowing the server only to make a halt at some of the queues (thus possibly forcing the server to keep switching for a while), e.g. at a queue which functions as a home base or at a queue where a new customer is most likely to arrive. The latter option may be recognized in the dynamic control of traffic lights. When there are no vehicles waiting, typically the main stream is given passage, until a waiting vehicle of a crossing stream is detected.

In many situations there are moreover significant cost involved in switching. In manufactur-

ing and maintenance applications e.g. the switch-over usually represents the change-over from one type of jobs to another, which may involve labour cost, material cost, or transportation cost. In such situations a potential saving in switching cost is an additional reason for letting the server stop switching when there are no customers present in the system.
Apart from the practical relevance, it is theoretically interesting to gain some insight into the effect of idling. In the first part of the present paper we therefore derive a pseudo-conservation law for a general model in which the server may be allowed to make a halt at a queue when there are no customers present in the system. A pseudo-conservation law provides a simple expression for a weighted sum of the mean waiting times. By its comparative simplicity a pseudo-conservation law is likely to provide some insight into the effect of idling, whereas the individual mean waiting times themselves involve expressions far too complicated to do so. Linked up with this, the determination of the individual mean waiting times, if at all possible, requires an intricate analysis and relies substantially on the features of the model under consideration, whereas the derivation of a pseudo-conservation law in fact only demands the calculation of mean working/idling times and only marginally leans on the characteristics of the model under consideration.
The option of idling especially goes hand in hand quite naturally with the globally gated service discipline, recently introduced by Boxma, Levy, & Yechiali [7]. The globally gated service discipline operates as follows. Suppose the server arrives at its home base. Then all the customers in the system are marked instantaneously and the server immediately starts a tour along the queues. During this tour exactly the marked customers are served. The service of customers that meanwhile arrive in the system is deferred until the next tour along the queues. Boxma, Weststrate, & Yechiali [9] propose the globally gated service discipline to be used by a repair crew, in charge of the maintenance activities at several installations. As indicated in [9], under the globally gated service discipline it does not make sense to start a tour along the queues when there are no customers present in the system. In the second part of the present paper we therefore direct the attention in particular to a globally gated polling system in which the server makes a halt at its home base when there are no customers present in the system. The globally gated service discipline then operates as follows. Suppose again the server arrives at its home base. If there are customers present in the system, they are all marked instantaneously and the server starts a tour along the queues, acting as described before. If there are no customers present in the system, the server remains idling at its home base, awaiting a new customer to arrive at one of the queues. As soon as a new customer arrives, it is marked instantaneously and the server starts a tour along the queues. During this tour only the newly arrived customer is served. The service of customers that meanwhile arrive in the system is again deferred until the next tour along the queues. As a justification of the dormant server policy we show that under the globally gated service discipline the waiting time at each of the queues is smaller (in the increasing convex ordering sense) than in the ordinary non-dormant server case.
The remainder of the paper is organized as follows. In section 2 we present a detailed model description. We derive a pseudo-conservation law for a general model in section 3 and use it in section 4 to compare the dormant and the non-dormant server case. We further address the question at which queues the server should make a halt to minimize the mean total amount of work in the system. Subsequently we direct the attention in particular to a globally gated polling system in which the server makes a halt at its home base when there are no customers

present in the system. For that system we derive an explicit expression for the LST of the cycle time distribution, for the LST of the waiting time distribution at each of the queues, and for the pgf of the joint queue length distribution at polling epochs in section 5, 6, and 7 respectively. The waiting time at each of the queues is shown to be smaller (in the increasing convex ordering sense) than in the ordinary non-dormant server case. In section 8 we present a conclusion.

2 MODEL DESCRIPTION

The model under consideration consists of n queues, $Q_1, \ldots, Q_n$, each of infinite capacity, attended to by a single server S. Customers arrive at the various queues according to independent Poisson processes. Customers arriving at Q_i will also be referred to as type-i customers. 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 having distribution $B_i(\cdot)$, with LST $\beta_i(\cdot)$, first moment β_i and second moment $\beta_i^{(2)}$, $i = 1, \ldots, n$. All service times are assumed to be independent. 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 visits the queues in a strictly cyclic order, $Q_1, \ldots, Q_n$. Moving from Q_i to Q_{i+1}, where $n+1$ is to be understood as 1, the server experiences a switch-over time, having distribution $S_i(\cdot)$, with LST $\sigma_i(\cdot)$, first moment s_i and second moment $s_i^{(2)}$, $i = 1, \ldots, n$. All switch-over times are assumed to be independent. The total switch-over time during a cycle has distribution $S(\cdot)$, with LST $\sigma(\cdot) := \prod_{i=1}^{n} \sigma_i(\cdot)$, first moment $s := \sum_{i=1}^{n} s_i$ and second moment $s^{(2)} := \sum_{i=1}^{n} s_i^{(2)} + \sum_{i=1}^{n} \sum_{j \neq i} s_i s_j$. The arrival process, the service process, and the switch-over process are assumed to be mutually independent.

As soon as the server arrives at Q_i, it starts serving type-i customers (possibly none), as prescribed by the service discipline. For now we do not specify the service disciplines at the various queues. We only demand the service disciplines at the various queues to be non-preemptive and work-conserving, i.e., no work is created or destroyed, cf. Boxma [4]. As soon as the server has finished serving type-i customers (possibly none), as prescribed by the service discipline, it departs from Q_i. However, we put in the proviso that the server may be allowed to make a halt at Q_i when there are no customers present in the system. The server then remains idling at Q_i, awaiting a new customer to arrive at one of the queues. For now we do not specify the additional criteria for deciding when to make a halt. In case of the 1-limited service discipline e.g., it seems to make sense only to make a halt at a queue when the server has not served any customer yet. As soon as a new customer arrives, the server resumes its activities. If the new customer arrives at Q_i and if the service discipline permits to do so, then the server immediately starts serving the newly arrived customer. Otherwise the server immediately starts switching to Q_{i+1}. The newly arrived customer then does not need to be served first, as, on its way to the newly arrived customer, the server may encounter other customers at other queues.

The case in which the server makes a halt at none of the queues when there are no customers present in the system, will also be referred to as the non-dormant server case. The case in

which the server makes a halt at all of the queues when there are no customers present in the system, will also be referred to as the pure dormant server case.
Finally some words on the ergodicity conditions. We claim - without proof - that the additional criteria for deciding when to make a halt do not affect the ergodicity conditions. If the ergodicity conditions are violated, then with probability 0 there are no customers present in the system, i.e., with probability 0 the opportunity to idle arises, so the additional criteria for deciding when to make a halt are irrelevant then. It therefore suffices to find the ergodicity conditions for the non-dormant server case. Obviously $\rho < 1$ is a necessary condition. Without pretending to be precise, we claim that $\rho < 1$ is also a sufficient condition for service disciplines, like exhaustive and gated, which do not impose any parametric (possibly probabilistic) restriction on the number of customers served during a visit, cf. Altman, Konstantopoulos, & Liu [2]. An additional condition for service disciplines, like 1-limited, which do impose such a parametric restriction, is that the mean number of customers that arrive during a cycle is smaller than the mean number of customers that are permitted to be served during a visit. Throughout the paper the ergodicity conditions are assumed to hold.

3 A PSEUDO-CONSERVATION LAW

As argued in the Introduction, a pseudo-conservation law is likely to provide some insight into the effect of idling. In the present section we therefore derive a pseudo-conservation law for the model under consideration. In section 4 we compare the dormant and the non-dormant server case on the basis of the pseudo-conservation law.
We first introduce some notation. Denote by π_i the probability that at an arbitrary epoch the server is idling at Q_i, $i = 1, \ldots, n$. If the server is not allowed to make a halt at Q_i, then of course $\pi_i = 0$. In general the probabilities π_i are not simply known. For the exhaustive and gated service discipline they can be determined along the lines of Eisenberg [12], as indicated in the Introduction. For the Bernoulli service discipline (comprising the exhaustive and 1-limited service discipline as extreme cases), they can be determined numerically using the power-series algorithm, cf. Blanc & Van der Mei [3]. The total probability that at an arbitrary epoch the server is idling is $\pi := \sum_{i=1}^{n} \pi_i$. Denote by $\mathbf{C}_i$ the cycle time with respect to Q_i, i.e., the time between two successive polling epochs at Q_i, $i = 1, \ldots, n$. Although the distribution of $\mathbf{C}_i$ in general depends on i, $\mathrm{E}\mathbf{C}_i$ obviously does not. Noting that the server is working a fraction ρ of the time and idling a fraction π of the time,

$$\mathrm{E}\mathbf{C}_i = \frac{s}{1-\rho-\pi}, \qquad i = 1, \ldots, n. \tag{3.1}$$

Denote by $\mathbf{U}_i$ the total time that the server is working at Q_i during a cycle, $i = 1, \ldots, n$. As ρ_i is the fraction of time that the server is working at Q_i, using (3.1),

$$\mathrm{E}\mathbf{U}_i = \frac{\rho_i s}{1-\rho-\pi}, \qquad i = 1, \ldots, n. \tag{3.2}$$

Denote by $\mathbf{I}_i$ the total time that the server is idling at Q_i during a cycle, $i = 1, \ldots, n$. As π_i is the fraction of time that the server is idling at Q_i, using (3.1),

$$\mathrm{E}\mathbf{I}_i = \frac{\pi_i s}{1-\rho-\pi}, \qquad i = 1, \ldots, n. \tag{3.3}$$

We now derive the pseudo-conservation law for the model under consideration. The approach is similar to the approach in Boxma & Groenendijk [5] for the ordinary non-dormant server case. It is easily verified that the model under consideration satisfies the assumptions mentioned in Boxma [4]. Hence, the following work decomposition property holds:

$$\mathbf{V} \stackrel{d}{=} \mathbf{V}_{M/G/1} + \mathbf{Y}, \tag{3.4}$$

where $\stackrel{d}{=}$ denotes equality in distribution and
$\mathbf{V}$:= amount of work in the system at an arbitrary epoch,
$\mathbf{V}_{M/G/1}$:= amount of work in the 'corresponding' system without switch-over times at an arbitrary epoch,
$\mathbf{Y}$:= amount of work in the system at an arbitrary epoch in a non-serving interval, i.e., a switching interval or an idling interval;
$\mathbf{V}_{M/G/1}$ and $\mathbf{Y}$ are independent.
The 'corresponding' system without switch-over times is a single server system with the same arrival process, service process, and service disciplines at the various queues, but without server interruptions from the switch-over process. A sample path comparison shows that the amount of work in the corresponding system without switch-over times does not depend on the service disciplines at the various queues.
Denote by $\mathbf{L}_i$ the number of waiting, i.e., not being served, type-i customers at an arbitrary epoch, $i = 1, \ldots, n$. As ρ_i is the probability that at an arbitrary epoch a type-i customer is being served, $i = 1, \ldots, n$,

$$\mathrm{E}\mathbf{V} = \sum_{i=1}^{n} \beta_i \mathrm{E}\mathbf{L}_i + \sum_{i=1}^{n} \rho_i \frac{\beta_i^{(2)}}{2\beta_i}. \tag{3.5}$$

Denote by $\mathbf{W}_i$ the waiting time of an arbitrary type-i customer, i.e., the time between its arrival and the start of its service, $i = 1, \ldots, n$. Using Little's law,

$$\mathrm{E}\mathbf{V} = \sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i + \frac{1}{2} \sum_{i=1}^{n} \lambda_i \beta_i^{(2)}. \tag{3.6}$$

Using the Pollaczek-Khintchine formula,

$$\mathrm{E}\mathbf{V}_{M/G/1} = \frac{\sum_{i=1}^{n} \lambda_i \beta_i^{(2)}}{2(1-\rho)}. \tag{3.7}$$

Combining (3.4), (3.6) and (3.7),

$$\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i = \rho \frac{\sum_{i=1}^{n} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \mathrm{E}\mathbf{Y}. \tag{3.8}$$

We now determine $\mathrm{E}\mathbf{Y}$, by distinguishing whether the server is switching or idling and conditioning on the type of switching interval. Denote by $\mathbf{Y}_i$ the total amount of work in the system at an arbitrary epoch in a switching interval from Q_i to Q_{i+1}, $i = 1, \ldots, n$. By definition there are no customers present in the system when the server is idling, in other words, the total amount of work in the system at an arbitrary epoch in an idling interval is zero.

Hence, with $\mathbf{EI} := \sum_{i=1}^{n} \mathbf{EI}_i$,

$$\mathbf{EY} = \sum_{i=1}^{n} \frac{s_i}{s + \mathbf{EI}} \mathbf{EY}_i. \tag{3.9}$$

$\mathbf{Y}_i$ is composed of two terms, viz.:
i. $\mathbf{Z}_i$, the total amount of work in the system at the beginning of the switch-over from Q_i to Q_{i+1};
ii. the total amount of work that arrives in the system between the beginning of the switch-over from Q_i to Q_{i+1} and the epoch under consideration;
so

$$\mathbf{EY}_i = \mathbf{EZ}_i + \rho \frac{s_i^{(2)}}{2s_i}, \qquad i = 1, \ldots, n. \tag{3.10}$$

We now fragment $\mathbf{Z}_i$ further into work that accumulated at different queues during different periods. Denote by $\mathbf{Z}_{ij}$ the amount of work at Q_j at the beginning of a switch-over from Q_i to Q_{i+1}, $i = 1, \ldots, n$, $j = 1, \ldots, n$. Then

$$\mathbf{EZ}_i = \sum_{j=1}^{n} \mathbf{EZ}_{ij}, \qquad i = 1, \ldots, n. \tag{3.11}$$

$\mathbf{Z}_{ij}$ is composed of two terms, viz.:
i. $\mathbf{Z}_{jj}$, the amount of work at Q_j at the beginning of the switch-over from Q_j to Q_{j+1};
ii. the amount of work that arrives at Q_j between the beginning of the switch-over from Q_j to Q_{j+1} and the beginning of the switch-over from Q_i to Q_{i+1};
so, using (3.2) and (3.3),

$$\mathbf{EZ}_{ij} = \mathbf{EZ}_{jj} + \rho_j \sum_{k=j+1}^{i} \left(s_{k-1} + \frac{(\rho_k + \pi_k)s}{1 - \rho - \pi} \right), \qquad i = 1, \ldots, n, i \neq j, \tag{3.12}$$

where the summation is to be interpreted cyclically.
Substituting (3.12) into (3.11),

$$\begin{aligned} \mathbf{EZ}_i &= \sum_{j=1}^{n} \mathbf{EZ}_{jj} + \sum_{j \neq i} \rho_j \sum_{k=j+1}^{i} \left(s_{k-1} + \frac{(\rho_k + \pi_k)s}{1 - \rho - \pi} \right) \\ &= \sum_{j=1}^{n} \mathbf{EZ}_{jj} + \sum_{j \neq i} \sum_{k=j}^{i-1} \rho_j s_k + \frac{s}{1 - \rho - \pi} \sum_{j=1}^{n} \sum_{k=1}^{j-1} \rho_j \rho_k + \frac{s}{1 - \rho - \pi} \sum_{j \neq i} \sum_{k=j+1}^{i} \rho_j \pi_k. \end{aligned} \tag{3.13}$$

Substituting (3.10) and (3.13) into (3.9), noting that $\frac{s}{s + \mathbf{EI}} = \frac{1 - \rho - \pi}{1 - \rho}$,

$$\begin{aligned} \mathbf{EY} = {} & \frac{s}{1 - \rho} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \rho_i \rho_j + \frac{\rho(1 - \rho - \pi)}{(1 - \rho)s} \left[\sum_{i=1}^{n} \sum_{j=1}^{i-1} s_i s_j + \frac{1}{2} \sum_{i=1}^{n} s_i^{(2)} \right] + \\ & \frac{1}{1 - \rho} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=j+1}^{i} s_i \rho_j \pi_k + \frac{1 - \rho - \pi}{1 - \rho} \sum_{i=1}^{n} \mathbf{EZ}_{ii} \end{aligned}$$

$$= \frac{s}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{n}\rho_i^2\right] + \frac{1-\rho-\pi}{1-\rho}\rho\frac{s^{(2)}}{2s} + \frac{1}{1-\rho}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=j+1}^{i} s_i\rho_j\pi_k + \frac{1-\rho-\pi}{1-\rho}\sum_{i=1}^{n}\mathrm{E}\mathbf{Z}_{ii}. \tag{3.14}$$

Substituting (3.14) into (3.8),

$$\sum_{i=1}^{n}\rho_i\mathrm{E}\mathbf{W}_i = \rho\frac{\sum_{i=1}^{n}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{s}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{n}\rho_i^2\right] + \frac{1-\rho-\pi}{1-\rho}\rho\frac{s^{(2)}}{2s} + \frac{1}{1-\rho}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=j+1}^{i} s_i\rho_j\pi_k + \frac{1-\rho-\pi}{1-\rho}\sum_{i=1}^{n}\mathrm{E}\mathbf{Z}_{ii}. \tag{3.15}$$

We may interpret the terms in (3.15) as follows. The term in the left-hand side is the mean amount of waiting work in the system at an arbitrary epoch. The first term in the right-hand side is the mean amount of waiting work in the corresponding system without switch-over times at an arbitrary epoch. The remaining terms in the right-hand side reflect the influence of the switch-over times. Together these terms constitute the mean total amount of work in the system at an arbitrary epoch in a non-serving interval, i.e., a switching interval or an idling interval. The last term in the right-hand side represents the mean amount of work at an arbitrary epoch in a non-serving interval that is left behind by the server at the various queues. Together the second, third, and fourth term in the right-hand side constitute the mean total amount of work at an arbitrary epoch in a non-serving interval that arrived at the various queues since the server left those queues. Separately the second, third, and fourth term represent the mean amount of work at an arbitrary epoch in a non-serving interval that arrived at the various queues during respectively the working, switching, and idling intervals since the server left those queues. These terms do not depend on the service disciplines at the various queues, at least as far as their global structure is concerned; the probabilities π_i that occur in these terms probably do depend on the service disciplines at the various queues. Apparently, at least as far as the global structure of (3.15) is concerned, the last term in the right-hand side completely captures the influence of the service disciplines at the various queues.

Remember that we determined the terms in the right-hand side of (3.15), except the first term, by distinguishing whether the server is switching or idling and conditioning on the type of switching interval. The second, third, and fifth term may in fact also be determined without conditioning on the type of switching interval, the fourth term does not allow such an alternative determination.

The last term in the right-hand side of (3.15), representing the mean amount of work that is left behind by the server at the various queues, still remains to be specified. Obviously $\mathrm{E}\mathbf{Z}_{ii}$ is influenced by the service discipline at Q_i. However, as a pleasing circumstance, $\mathrm{E}\mathbf{Z}_{ii}$ is in fact influenced by the service discipline at Q_i *only*, i.e., not by the service discipline at Q_j, $j \neq i$. On the basis of the service discipline at Q_i only, we can split $\mathrm{E}\mathbf{Z}_{ii}$ into work that arrived during working, switching, and idling intervals, whose means do not depend on the service disciplines at the various queues, cf. (3.2) and (3.3).

We now determine $\mathrm{E}\mathbf{Z}_{ii}$ for the exhaustive (I), gated (II), 1-limited (III), and globally gated (IV) service discipline. We need to distinguish whether the server, when idling at Q_i, has already served a customer during its visit to Q_i or not. Denote by π_i' the probability that at an arbitrary epoch the server is idling at Q_i and has not served any customer yet, $i = 1, \ldots, n$. Denote by π_i'' the probability that at an arbitrary epoch the server is idling at Q_i and has already served a customer, $i = 1, \ldots, n$. Neither π_i' nor π_i'' are simply known, but of course $\pi_i = \pi_i' + \pi_i''$. Let E, G, L, and GG represent respectively the set of queues where the exhaustive, gated, 1-limited, and globally gated service discipline is used.

I. Exhaustive: S serves type-i customers until Q_i is empty. So,

$$\mathrm{E}\mathbf{Z}_{ii} = 0, \qquad i \in E. \tag{3.16}$$

II. Gated: S serves exactly those type-i customers present at its arrival at Q_i. A customer that arrives when S is idling at Q_i and has not served any customer yet is also served. So, using (3.2) and (3.3),

$$\mathrm{E}\mathbf{Z}_{ii} = \frac{\rho_i(\rho_i + \pi_i'')s}{1-\rho-\pi}, \qquad i \in G. \tag{3.17}$$

III. 1-Limited: S serves 1 type-i customer, provided there is a customer present at its arrival at Q_i. A customer that arrives when S is idling at Q_i and has not served any customer yet is also served. Thus S leaves behind the amount of work that arrives during the waiting time of the customer that is possibly served and during its visit to Q_i, but not during the possible idling period before S serves a customer. It follows from (3.1) that on average $\dfrac{\lambda_i s}{1-\rho-\pi}$ customers are served during a visit to Q_i. In particular, under the 1-limited service discipline, with probability $\dfrac{\lambda_i s}{1-\rho-\pi}$ a customer is served during a visit to Q_i. So, using (3.2) and (3.3),

$$\mathrm{E}\mathbf{Z}_{ii} = \rho_i \left(\frac{\lambda_i s}{1-\rho-\pi} \mathrm{E}\mathbf{W}_i + \frac{(\rho_i + \pi_i'')s}{1-\rho-\pi} \right), \qquad i \in L. \tag{3.18}$$

As claimed in section 2, an additional ergodicity condition for the 1-limited service discipline is that the mean number of customers that arrive during a cycle is smaller than the mean number of customers that are permitted to be served during a visit: $\dfrac{\lambda_i s}{1-\rho-\pi} < 1$. As further claimed in section 2, the ergodicity conditions do not differ from the ergodicity conditions for the non-dormant server case: $\dfrac{\lambda_i s}{1-\rho} < 1$. In other words, $\dfrac{\lambda_i s}{1-\rho} < 1$ should guarantee $\dfrac{\lambda_i s}{1-\rho-\pi} < 1$. Notice that the latter condition yields the bound $\pi < 1 - \rho - \lambda_i s$, $i \in L$, which may improve significantly upon the trivial bound $\pi < 1 - \rho$.

IV. Globally gated: S serves exactly those type-i customers present at its most recent arrival at Q_1. A customer that arrives when S is idling at Q_1 and has not served any customer yet is also served. So, using (3.2) and (3.3),

$$\mathrm{E}\mathbf{Z}_{ii} = \rho_i \left(\frac{(\rho_1 + \pi_1'')s}{1-\rho-\pi} + \sum_{j=2}^{i} \left(s_{j-1} + \frac{(\rho_j + \pi_j)s}{1-\rho-\pi} \right) \right), \qquad i \in GG. \tag{3.19}$$

Substituting (3.16), (3.17), (3.18), and (3.19) into (3.15), we obtain the following pseudo-conservation law:

Theorem 3.1
The mean waiting times in the model under consideration satisfy the following relationship:

$$\sum_{i\in E,G,GG} \rho_i \mathrm{EW}_i + \sum_{i\in L} \rho_i \left(1 - \frac{\lambda_i s}{1-\rho}\right) \mathrm{EW}_i = \rho \frac{\sum_{i=1}^{n} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \tag{3.20}$$

$$\frac{s}{2(1-\rho)} \left[\rho^2 - \sum_{i\in E} \rho_i^2 + \sum_{i\in G,L,GG} \rho_i^2 + 2 \sum_{i\in GG} \sum_{j=1}^{i-1} \rho_i \rho_j\right] + \frac{1-\rho-\pi}{1-\rho} \left[\rho \frac{s^{(2)}}{2s} + \sum_{i\in GG} \sum_{j=1}^{i-1} \rho_i s_j\right] +$$

$$\frac{1}{1-\rho} \left[\sum_{i=1}^{n} \sum_{j\neq i} \sum_{k=j+1}^{i} s_i \rho_j \pi_k + s \sum_{i\in G,L} \rho_i \pi_i'' + s \sum_{i\in GG} \rho_i (\pi_1'' + \sum_{j=2}^{i} \pi_j)\right].$$

Elaborating on the explanation of the terms in (3.15), we may interpret the terms in (3.20) as follows. The terms in the left-hand side, excluding the term $\sum_{i\in L} \rho_i \frac{\lambda_i s}{1-\rho} \mathrm{EW}_i$, together constitute the mean amount of waiting work in the system at an arbitrary epoch. The first term in the right-hand side is the mean amount of waiting work in the corresponding system without switch-over times at an arbitrary epoch. The remaining terms in the right-hand side and the term $\sum_{i\in L} \rho_i \frac{\lambda_i s}{1-\rho} \mathrm{EW}_i$ together constitute the mean total amount of work in the system at an arbitrary epoch in a non-serving interval, i.e., a switching interval or an idling interval. Separately the second, third, and fourth term represent the mean amount of work at an arbitrary epoch in a non-serving interval that arrived at the various queues during respectively working, switching, and idling intervals, excluding the work that arrived at 1-limited queues before the server polled those queues for the last time.

Remark 3.1 For $\pi_i = 0$, $i = 1, \ldots, n$, (3.20) reduces to the pseudo-conservation law for the ordinary non-dormant server case, cf. Boxma & Groenendijk [5].

□

Remark 3.2 It is easily verified that the mean waiting times in the model of Eisenberg [11] with the exhaustive service discipline at both queues indeed satisfy (3.20) with $n = 2$, $E = \{1, 2\}$.

□

Remark 3.3 Notice that the pseudo-conservation law is not as powerful in the dormant server case as in the ordinary non-dormant server case. In the non-dormant server case the pseudo-conservation law provides a simple expression for a weighted sum of the mean waiting times, although the individual mean waiting times themselves in general involve quite complicated expressions. In the dormant server case the pseudo-conservation law still provides a simple expression for a weighted sum of the mean waiting times, at least as far as its global structure is concerned; the probabilities π_i that occur probably do hide rather

cumbersome expressions. Linked up with this, in the dormant server case, to obtain the pseudo-conservation law explicitly, we are committed to an intricate analysis to determine the probabilities π_i, if at all possible. For the exhaustive and gated service discipline they can be determined along the lines of Eisenberg [12], as indicated in the Introduction. For the globally gated service discipline, mathematically a most tractable service discipline, there is no obstacle to determining the probabilities π_i either. For the 1-limited service discipline there is no method available for determining the probabilities π_i analytically. However, they can be determined numerically using the power-series algorithm, cf. Blanc & Van der Mei [3]. □

Remark 3.4 Like in the ordinary non-dormant server case, cf. Groenendijk [13], a pseudo-conservation law is useful for supporting approximations for the mean waiting times and for finding the exact mean waiting times in a symmetrical system in a simple manner. Various approximations for the mean waiting times are conceivable, but we do not pursue this matter here any further. □

Remark 3.5 In the present section we derived a pseudo-conservation law for a model with cyclic polling and single Poisson arrivals. Without seriously complicating the above analysis, cyclic polling may be generalized to polling guided by a table, cf. [6], or Markovian polling, i.e., the server visits the queues guided by a Markov chain with state space $\{1,\ldots,n\}$, cf. [8], and single Poisson arrivals may be generalized to batch Poisson arrivals, cf. Boxma [4]. □

4 Comparison between the dormant and the non-dormant server case

In section 3 we derived a pseudo-conservation law for the model under consideration. In the present section we compare the dormant and the non-dormant server case on the basis of the pseudo-conservation law, to gain some insight into the effect of idling. Specifically we compare the mean waiting times in a symmetrical system in the pure dormant and the non-dormant server case. We further address the question at which queues the server should make a halt to minimize the mean total amount of work in the system.

Let us label the waiting times in the dormant and the non-dormant server case with a hat and a tilde respectively. From (3.20),

$$\sum_{i\in E,G,GG} \rho_i(\mathrm{E}\hat{\mathbf{W}}_i - \mathrm{E}\tilde{\mathbf{W}}_i) + \sum_{i\in L} \rho_i \left(1 - \frac{\lambda_i s}{1-\rho}\right)(\mathrm{E}\hat{\mathbf{W}}_i - \mathrm{E}\tilde{\mathbf{W}}_i) = \tag{4.1}$$

$$-\frac{\pi}{1-\rho}\left[\rho\frac{s^{(2)}}{2s} + \sum_{i\in GG}\sum_{j=1}^{i-1}\rho_i s_j\right] +$$

$$\frac{1}{1-\rho}\left[\sum_{i=1}^{n}\sum_{i\neq j}\sum_{k=j+1}^{i} s_i\rho_j\pi_k + s\sum_{i\in G,L}\rho_i\pi_i'' + s\sum_{i\in GG}\rho_i(\pi_1'' + \sum_{j=2}^{i}\pi_j)\right] =$$

$$-\frac{\rho\pi}{1-\rho}\left\{\left[\frac{s^{(2)}}{2s}+\sum_{i\in GG}\sum_{j=1}^{i-1}\frac{\rho_i}{\rho}s_j\right]-\right.$$

$$\left.\left[\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=j+1}^{i}s_i\frac{\rho_j}{\rho}\frac{\pi_k}{\pi}+s\sum_{i\in G,L}\frac{\rho_i}{\rho}\frac{\pi_i''}{\pi}+s\sum_{i\in GG}\frac{\rho_i}{\rho}\left(\frac{\pi_1''}{\pi}+\sum_{j=2}^{i}\frac{\pi_j}{\pi}\right)\right]\right\},$$

where π of course refers to the dormant server case. The interpretation of the terms in the left-hand side and in the first form of the right-hand side of (4.1) follows immediately from the interpretation of the terms in (3.20). The terms in the left-hand side, excluding the term $\sum_{i\in L}\rho_i\frac{\lambda_i s}{1-\rho}(\mathrm{E}\hat{\mathbf{W}}_i-\mathrm{E}\tilde{\mathbf{W}}_i)$, together constitute the difference in the mean total amount of work in the system at an arbitrary epoch between the dormant and the non-dormant server case.

The first term in the first form of the right-hand side represents the difference in the mean amount of work at an arbitrary epoch in a non-serving interval that arrived during *switching* intervals (excluding the work that arrived at 1-limited queues during switching intervals before the server polled those queues for the last time). The first term itself is the product of two terms. The term inside the square brackets represents the mean amount of work at an arbitrary epoch in a switching interval that arrived during switching intervals (excluding the work that arrived at 1-limited queues during switching intervals before the server polled those queues for the last time). Notice that these quantities are the same in the dormant and the non-dormant server case. The term in front is the probability that in the dormant server case an arbitrary epoch in a non-serving interval concerns an idling epoch rather than a switching epoch. The first term is the product of these two terms, as the amount of work at an arbitrary idling epoch in the dormant server case is by definition zero.

The second term in the first form of the right-hand side represents the difference in the mean amount of work at an arbitrary epoch in a non-serving interval that arrived during *idling* intervals. The second term is of course just the mean amount of work at an arbitrary epoch in a non-serving interval that arrived during idling intervals in the dormant server case, as in the non-dormant server case there are by definition no idling intervals.

The interpretation of the terms in the second form of the right-hand side of (4.1) requires a somewhat different point of view. Define $f(i,j,k)$, $g(i,j,k)$, and $h(i,j,k)$ as the number of times the server needs to switch from Q_i to Q_{i+1} when currently in an empty system respectively switching from Q_k to Q_{k+1}, idling at Q_k not having served any customer yet, and idling at Q_k having already served a customer, before it can perform work currently arriving at Q_j, *taking into account the service discipline at* Q_j. Then the first and the second term in the first form of the right-hand side may be written as

$$\frac{\pi}{1-\rho}\sum_{i=1}^{n}\frac{s_i}{s}\sum_{j=1}^{n}\rho_j\left(\frac{s_i^{(2)}}{2s_i}+\sum_{k=1}^{n}s_k f(i,j,k)\right)=\frac{\pi\rho}{1-\rho}\sum_{i=1}^{n}\frac{s_i}{s}\sum_{j=1}^{n}\frac{\rho_j}{\rho}\left(\frac{s_i^{(2)}}{2s_i}+\sum_{k=1}^{n}s_k f(k,j,i)\right)$$

and

$$\frac{1-\rho-\pi}{1-\rho}\sum_{i=1}^{n}\frac{s_i}{s}\sum_{j=1}^{n}\rho_j\sum_{k=1}^{n}\left(\frac{\pi_k' s}{1-\rho-\pi}g(i,j,k)+\frac{\pi_k'' s}{1-\rho-\pi}h(i,j,k)\right)=$$

$$\frac{\pi\rho}{1-\rho}\left(\sum_{i=1}^{n}\frac{\pi_i'}{\pi}\sum_{j=1}^{n}\frac{\rho_j}{\rho}\sum_{k=1}^{n}s_k g(k,j,i)+\sum_{i=1}^{n}\frac{\pi_i''}{\pi}\sum_{j=1}^{n}\frac{\rho_j}{\rho}\sum_{k=1}^{n}s_k h(k,j,i)\right)$$

respectively. As arbitrary work concerns work at Q_j with probability $\frac{\rho_j}{\rho}$, we may interpret

$$\sum_{i=1}^{n}\frac{s_i}{s}\sum_{j=1}^{n}\frac{\rho_j}{\rho}\left(\frac{s_i^{(2)}}{2s_i}+\sum_{k=1}^{n}s_k f(k,j,i)\right)$$

and

$$\sum_{i=1}^{n}\frac{\pi_i'}{\pi}\sum_{j=1}^{n}\frac{\rho_j}{\rho}\sum_{k=1}^{n}s_k g(k,j,i)+\sum_{i=1}^{n}\frac{\pi_i''}{\pi}\sum_{j=1}^{n}\frac{\rho_j}{\rho}\sum_{k=1}^{n}s_k h(k,j,i)$$

as the mean total switch-over time to be incurred by the server when currently in an empty system switching and idling respectively, before it can perform arbitrary currently arriving work, *taking into account the various service disciplines.* For brevity let us refer to these quantities as the mean distance to arbitrary work for a switching and an idling server respectively. Concluding, the difference in the mean total amount of work between the dormant and the non-dormant server case is just the difference in the mean distance to arbitrary work between a switching and an idling server, preceded by the multiplier $\frac{\pi\rho}{1-\rho}$.

We now compare the mean waiting times in a symmetrical system in the pure dormant and the non-dormant server case for the exhaustive (I), gated (II), and 1-limited (III) service discipline. To do so we take in (4.1) $\rho_i=\rho/n$, $s_i=s/n$, $\pi_i=\pi/n$, $\mathrm{E}\hat{\mathbf{W}}_i=\mathrm{E}\hat{\mathbf{W}}$, $\mathrm{E}\tilde{\mathbf{W}}_i=\mathrm{E}\tilde{\mathbf{W}}$ and for the 1-limited service discipline in addition $\lambda_i=\lambda/n$, $\pi_i''=\pi''/n$.
We can not compare the mean waiting times in a symmetrical system for the globally gated service discipline. Even when the Q_i are all identical, the asymmetrical nature of the globally gated service discipline will preclude that the mean waiting times $\mathrm{E}\mathbf{W}_i$ and the probabilities π_i are all identical.
I. Exhaustive:

$$\mathrm{E}\hat{\mathbf{W}}-\mathrm{E}\tilde{\mathbf{W}}=\frac{\pi s}{2(1-\rho)}\left[\frac{n-1}{n}-\frac{s^{(2)}}{s^2}\right]\leq-\frac{\pi s}{2n(1-\rho)}\leq 0. \tag{4.2}$$

II. Gated:

$$\mathrm{E}\hat{\mathbf{W}}-\mathrm{E}\tilde{\mathbf{W}}=\frac{s}{2(1-\rho)}\left[\pi\left(\frac{n-1}{n}-\frac{s^{(2)}}{s^2}\right)+\frac{2\pi''}{n}\right]\leq\frac{\pi s}{2n(1-\rho)}\leq\frac{s}{2n}. \tag{4.3}$$

III. 1-Limited:

$$\mathrm{E}\hat{\mathbf{W}}-\mathrm{E}\tilde{\mathbf{W}}=\frac{s}{2(1-\rho-\frac{\lambda s}{n})}\left[\pi\left(\frac{n-1}{n}-\frac{s^{(2)}}{s^2}\right)+\frac{2\pi''}{n}\right]\leq\frac{\pi s}{2(n(1-\rho)-\lambda s)}\leq\frac{s}{2n}. \tag{4.4}$$

For the exhaustive service discipline always $E\hat{W} \leq E\tilde{W}$, which agrees with the result of Liu, Nain, & Towsley [16] that in a symmetrical system the server should remain idling at a queue when the entire system is empty to minimize the total amount of work. Also $E\hat{W} \leq E\tilde{W}$ for the gated service discipline when the server is only allowed to make a halt at a queue when it has not served any customer yet, i.e., when $\pi'' = 0$. When the server is also allowed to make a halt at a queue when it has already served a customer, i.e., when $\pi'' = \pi$, $E\hat{W} \leq E\tilde{W}$ ($E\hat{W} \geq E\tilde{W}$) for the gated service discipline, iff the coefficient of variation of the total switch-over time is larger (smaller) than $1 + 1/n$. In particular $E\hat{W} = E\tilde{W}$ when the total switch-over time is Erlang-n distributed, so when the individual switch-over times are exponentially distributed. Because of the memoryless property of the exponential distribution customers then indeed do not observe any difference between the dormant and the non-dormant server case. Similar remarks hold for the 1-limited service discipline.

We finally address the question at which queues the server should make a halt to minimize the mean total amount of work in the system. Liu, Nain, & Towsley [16] identify polling policies, allowing idling as a possible action, that stochastically minimize the total amount of work in the system at an arbitrary epoch. They show that optimal policies are exhaustive, greedy, and in a symmetrical system in addition patient, i.e., the server should neither switch nor idle when at a non-empty queue and in a symmetrical system the server should remain idling at a queue when the entire system is empty. As a non-exhaustive policy can not minimize the total amount of work, we assume in the sequel the service discipline to be exhaustive.
In asymmetrical systems the total amount of work is not always minimal in the pure dormant server case. In some asymmetrical systems the total amount of work is even in the non-dormant server case smaller. Consider e.g. a system with $n = 2$, $\lambda_1 = \infty$, $\beta_1 = 0$, $s_1^{(2)} = s_1^2$, $s_2 = 0$. As far as the amount of work is concerned, such a system corresponds in the pure dormant and the non-dormant server case to an ordinary $M/G/1$ queue with set-up times and multiple vacations, respectively, of length s_1. In the latter case the amount of work is of course smaller. However, in not a single system the total amount of work is minimal in the non-dormant server case, as we will prove now. In case the server makes a halt only at Q_h (4.1) reduces to

$$\sum_{i=1}^{n} \rho_i (E\hat{W}_i - E\tilde{W}_i) = -\frac{\pi_h}{1-\rho} \left[\rho \frac{s^{(2)}}{2s} - \sum_{i \neq h-1} \sum_{j=i+1}^{h-1} s_i \rho_j \right].$$

So it suffices to show that there is at least one Q_h such that $\sum_{i \neq h-1} \sum_{j=i+1}^{h-1} s_i \rho_j \leq \rho \frac{s^{(2)}}{2s}$. Now $\sum_{h=1}^{n} \sum_{i \neq h-1} \sum_{j=i+1}^{h-1} s_i \rho_j = \frac{1}{2} n \rho s - \sum_{i=1}^{n} s_i \rho_i$ implies $\min_{1 \leq h \leq n} \sum_{i \neq h-1} \sum_{j=i+1}^{h-1} s_i \rho_j \leq \frac{1}{2} \rho s \leq \rho \frac{s^{(2)}}{2s}$. So there is indeed at least one such Q_h.

We now *know* that making a halt only at Q_h is beneficial, iff $\sum_{i \neq h-1} \sum_{j=i+1}^{h-1} s_i \rho_j \leq \rho \frac{s^{(2)}}{2s}$. In addition we now *assume* that Q_h belongs to the set of queues at which the server should make a halt, when making a halt only at Q_h is beneficial. In other words, we propose

to let the server make a halt at Q_h, iff $\sum_{i\neq h-1}\sum_{j=i+1}^{h-1} s_i\rho_j \leq \rho\frac{s^{(2)}}{2s}$. In accordance with the interpretation of the second form of (4.1), the criterion is seen to select queues Q_h, such that the mean distance from the server to arbitrary work is smaller when idling at Q_h than when switching. Written as $\sum_{i\neq h-1}\sum_{j=i+1}^{h-1} \frac{s_i}{s}\frac{\rho_j}{\rho} \leq \frac{s^{(2)}}{2s^2}$, the criterion is seen to suggest idling at queues Q_h preceded by queues with relatively light traffic and large switch-over times and followed by queues with relatively heavy traffic and small switch-over times, and to suggest idling more strongly accordingly as the variability of the total switch-over time is larger. Notice that in a symmetrical system the inequality reduces to $1 - 1/n \leq \frac{s^{(2)}}{s^2}$, which always holds, cf. (4.2).

Remark 4.1 The problem at which queues the server should make a halt to minimize the mean total amount of work in the system, may be formulated as a semi-Markov decision problem, cf. Tijms [22] p. 200. The decision epochs are the epochs of a visit completion; the possible decisions (actions) are either switching or idling when the visit completion leaves the entire system empty and only switching otherwise. The states are $(i, l_1, \ldots, l_n)$, where i is the queue at which the visit completion occurs and l_j is the number of waiting customers at Q_j, $j = 1, \ldots, n$. The crucial observation is that the system under consideration satisfies the following Markovian property: given the state at some decision epoch and the decision (action) chosen, the evolution after that decision epoch does not depend on the evolution before that decision epoch.
A semi-Markov decision problem formulates the problem of finding a strategy that minimizes the mean total cost per unit of time. A strategy prescribes here at which queues the server should make a halt. The mean total cost per unit of time may be related here to the mean total amount of work in the system, when we define the cost appropriately. If c_i represents the waiting cost per unit of time of an arbitrary type-i customer, $i = 1, \ldots, n$, then the mean total cost per unit of time equal $\sum_{i=1}^{n} c_i\lambda_i \mathrm{E}\mathbf{W}_i$. When we take $c_i = \beta_i$, $i = 1, \ldots, n$, the mean total cost per unit of time equal $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$, the mean amount of waiting work in the system. Minimizing the mean amount of waiting work and minimizing the mean total amount of work are equivalent, as the difference, the mean amount of work in service, always equals $\frac{1}{2}\sum_{i=1}^{n} \lambda_i\beta_i^{(2)}$, cf. (3.6).
It is straightforward, when action a is chosen in the current state h, to calculate $p(a, h, k)$, the probability that at the next decision epoch the state will be k, $t(a, h)$, the expected time until the next decision epoch, and $c(a, h)$, the expected cost incurred until the next decision epoch. The resulting semi-Markov decision problem may be solved numerically by truncating the state space.

□

5 THE CYCLE TIME IN THE GLOBALLY GATED POLLING SYSTEM

As argued in the Introduction, the option of idling especially goes hand in hand quite naturally with the globally gated service discipline. From the present section on we therefore direct the attention in particular to a globally gated polling system in which the server makes a halt at its home base when there are no customers present in the system. We first present a specific model description. Suppose the server is just about to visit Q_1. If there are customers present in the system, they are all marked instantaneously and the server immediately starts visiting $Q_1, \ldots, Q_n$. During the coming cycle exactly those marked customers are served. At each queue customers are served in order of arrival. The service of customers that meanwhile arrive in the system is deferred until the next cycle. If there are no customers present in the system, the server remains idling at Q_1, awaiting a new customer to arrive at one of the queues. As soon as a new customer arrives, it is marked instantaneously and the server starts visiting $Q_1, \ldots, Q_n$. During the coming cycle only the newly arrived customer is served. Again, customers that meanwhile arrive in the system are served during the next cycle. During the cycle the server is not allowed to make a halt at a queue when the completion of a service leaves the system empty. In other words, the server is only allowed to make a halt when the system is empty at the beginning of a visit to Q_1.

Remark 5.1 For $n = 1$ the model under consideration reduces to a gated vacation model with single vacations, whereas the model in the non-dormant server case corresponds to a gated vacation model with multiple vacations, cf. Takagi [19]. The latter model has been analyzed in detail by Takine & Hasegawa [21].

□

In the remainder of the present section we relate the cycle time distribution to the joint queue length distribution at the beginning of a cycle and at the beginning of a subsequent cycle. The approach is similar to the approach in Boxma, Levy, & Yechiali [7] for the ordinary non-dormant server case. Assuming an equilibrium distribution, we obtain a functional equation for the pgf of the joint queue length distribution at the beginning of a cycle and for the LST of the cycle time distribution. The latter functional equation is solved explicitly. The cycle time distribution will turn out to play a crucial role in the analysis of the waiting time distribution and the joint queue length distribution at polling epochs in section 6 and 7 respectively.
We first introduce some notation. Denote by $\mathbf{C}^{(m)}$ the length of the m-th cycle, i.e., the time between the start of the m-th visit to Q_1 and the start of the $(m+1)$-th visit to Q_1, $m = 1, 2, \ldots$. Denote by $\mathbf{I}^{(m)}$ the length of the m-th idling period, i.e., the m-th idling time at Q_1 (possibly zero), $m = 1, 2, \ldots$. Denote by $\mathbf{B}^{(m)}$ the length of the m-th *restricted* cycle, i.e., the m-th cycle time minus the m-th idling time, $m = 1, 2, \ldots$. Let $\alpha_m(\zeta, \omega) := \mathrm{E}(\mathrm{e}^{-\zeta \mathbf{I}^{(m)} - \omega \mathbf{B}^{(m)}})$ for $\mathrm{Re}\zeta \geq 0$, $\mathrm{Re}\omega \geq 0$, $m = 1, 2, \ldots$. Let $\gamma_m(\omega) := \mathrm{E}(\mathrm{e}^{-\omega \mathbf{B}^{(m)}})$ for $\mathrm{Re}\omega \geq 0$, $m = 1, 2, \ldots$. Denote by $\mathbf{X}_i^{(m)}$ the number of customers present at queue i at the beginning of the m-th cycle, $i = 1, \ldots, n, m = 1, 2, \ldots$. Let $\xi_m(z_1, \ldots, z_n) := \mathrm{E}(z_1^{\mathbf{X}_1^{(m)}} \ldots z_n^{\mathbf{X}_n^{(m)}})$ for $|z_i| \leq 1, i = 1, \ldots, n, m = 1, 2, \ldots$.
By the very nature of the globally gated service discipline $\mathbf{I}^{(m)}$, $\mathbf{B}^{(m)}$, $\mathbf{X}_i^{(m)}$, and $\mathbf{X}_i^{(m+1)}$, $i = 1, \ldots, n$, are related as follows. On the one hand $\mathbf{X}_i^{(m)}$ equals the number of customers

that are served at Q_i during the m-th restricted cycle, $i = 1, \ldots, n$, unless $(\mathbf{X}_1^{(m)}, \ldots, \mathbf{X}_n^{(m)}) = (0, \ldots, 0)$, i.e., there are no customers present at the beginning of the m-th cycle. In that case the server remains idling at Q_1 for a period, which is negative exponentially distributed with parameter λ, until a customer arrives at one of the queues; such an arrival occurs at Q_i with probability λ_i/λ, $i = 1, \ldots, n$. So

$$\mathrm{E}(\mathrm{e}^{-\zeta \mathbf{I}^{(m)} - \omega \mathbf{B}^{(m)}} \mid \mathbf{X}_1^{(m)}, \ldots, \mathbf{X}_n^{(m)}) =$$

$$\sigma(\omega) \left[\prod_{i=1}^{n} \beta_i(\omega)^{\mathbf{X}_i^{(m)}} - \left(1 - \frac{\lambda}{\lambda + \zeta} \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \beta_i(\omega) \right) \mathrm{I}_{\{\mathbf{X}^{(m)} = 0\}} \right], \tag{5.1}$$

$$\mathrm{Re}\, \zeta \geq 0, \mathrm{Re}\, \omega \geq 0, m = 1, 2, \ldots,$$

where $\mathrm{I}_{\{\mathbf{X}^{(m)}=0\}}$ denotes the indicator function of the event $(\mathbf{X}_1^{(m)}, \ldots, \mathbf{X}_n^{(m)}) = (0, \ldots, 0)$. Unconditioning (5.1),

$$\alpha_m(\zeta, \omega) = \sigma(\omega) \left[\xi_m(\beta_1(\omega), \ldots, \beta_n(\omega)) - \left(1 - \frac{\lambda}{\lambda + \zeta} \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \beta_i(\omega) \right) \xi_m(0, \ldots, 0) \right], \tag{5.2}$$

$$\mathrm{Re}\, \zeta \geq 0, \mathrm{Re}\, \omega \geq 0, m = 1, 2, \ldots .$$

On the other hand $\mathbf{X}_i^{(m+1)}$ equals the number of customers that arrive at Q_i during the m-th restricted cycle, $i = 1, \ldots, n, m = 1, 2, \ldots$. So

$$\mathrm{E}(z_1^{\mathbf{X}_1^{(m+1)}} \ldots z_n^{\mathbf{X}_n^{(m+1)}} \mid \mathbf{B}^{(m)} = t) = \mathrm{e}^{-\sum_{i=1}^{n} \lambda_i (1 - z_i) t}, \tag{5.3}$$

$$t \geq 0, \mid z_i \mid \leq 1, i = 1, \ldots, n, m = 1, 2, \ldots .$$

Define $\epsilon(z) := \sum_{i=1}^{n} \lambda_i (1 - z_i)$ for $\mid z_i \mid \leq 1, i = 1, \ldots, n$. Unconditioning (5.3),

$$\xi_{m+1}(z_1, \ldots, z_n) = \gamma_m(\epsilon(z)), \qquad \mid z_i \mid \leq 1, i = 1, \ldots, n, m = 1, 2, \ldots . \tag{5.4}$$

Denote by $\mathbf{C}, \mathbf{I}, \mathbf{B}$, and $\mathbf{X}_i$ stochastic variables with the limiting distribution for $m \to \infty$ of, respectively, $\mathbf{C}^{(m)}, \mathbf{I}^{(m)}, \mathbf{B}^{(m)}$, and $\mathbf{X}_i^{(m)}$, $i = 1, \ldots, n$. Let $\alpha(\zeta, \omega) := \mathrm{E}(\mathrm{e}^{-\zeta \mathbf{I} - \omega \mathbf{B}})$ for $\mathrm{Re}\, \zeta \geq 0$, $\mathrm{Re}\, \omega \geq 0$. Let $\gamma(\omega) := \mathrm{E}(\mathrm{e}^{-\omega \mathbf{B}})$ for $\mathrm{Re}\, \omega \geq 0$. Let $\xi(z_1, \ldots, z_n) := \mathrm{E}(z_1^{\mathbf{X}_1}, \ldots, z_n^{\mathbf{X}_n})$ for $\mid z_i \mid \leq 1, i = 1, \ldots, n$.
From (5.2) and (5.4),

$$\alpha(\zeta, \omega) = \sigma(\omega) \left[\xi(\beta_1(\omega), \ldots, \beta_n(\omega)) - \left(1 - \frac{\lambda}{\lambda + \zeta} \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \beta_i(\omega) \right) \xi(0, \ldots, 0) \right], \tag{5.5}$$

$$\mathrm{Re}\, \zeta \geq 0, \mathrm{Re}\, \omega \geq 0,$$

and

$$\xi(z_1, \ldots, z_n) = \gamma(\epsilon(z)), \qquad \mid z_i \mid \leq 1, i = 1, \ldots, n\,. \tag{5.6}$$

Combining (5.5) and (5.6),

$$\alpha(\zeta,\omega) = \sigma(\omega)\left[\gamma(\sum_{i=1}^{n}\lambda_i(1-\beta_i(\omega))) - \gamma(\lambda)\left(1-\frac{\lambda}{\lambda+\zeta}\sum_{i=1}^{n}\frac{\lambda_i}{\lambda}\beta_i(\omega)\right)\right], \tag{5.7}$$

$$\mathrm{Re}\,\zeta \geq 0, \mathrm{Re}\,\omega \geq 0, \mid z_i \mid \leq 1, i = 1,\ldots,n,$$

and

$$\xi(z_1,\ldots,z_n) =$$

$$\sigma(\epsilon(z))\left[\xi(\beta_1(\epsilon(z)),\ldots,\beta_n(\epsilon(z))) - \left(1-\sum_{i=1}^{n}\frac{\lambda_i}{\lambda}\beta_i(\epsilon(z))\right)\xi(0,\ldots,0)\right], \tag{5.8}$$

$$\mid z_i \mid \leq 1, i = 1,\ldots,n.$$

Remark 5.2 The joint queue length process at the beginning of a cycle, $\{(\mathbf{X}_1^{(m)},\ldots,\mathbf{X}_n^{(m)})\}$, $m = 1,2,\ldots$, in fact constitutes a multitype branching process with state-dependent immigration, cf. Resing [17]. The crucial observation is that the globally gated service discipline satisfies the following property: if there are x_i customers present at Q_i at the beginning of a cycle, then each of these x_i customers will be 'effectively replaced' in an i.i.d. manner by a random population having pgf $\beta_i(\epsilon(z))$. Adopting the terminology of the theory of multitype branching processes, the offspring generating functions are given by $f_i(z_1,\ldots,z_n) = \beta_i(\epsilon(z))$, $i = 1,\ldots,n$, the immigration generating function for the non-zero states is given by $g(z_1,\ldots,z_n) = \sigma(\epsilon(z))$, and the immigration generating function for the zero state is given by $g(z_1,\ldots,z_n)h(z_1,\ldots,z_n)$ with $h(z_1,\ldots,z_n) = \sum_{i=1}^{n}\frac{\lambda_i}{\lambda}f_i(z)$. From the theory of multitype branching processes we have

$$\xi(z_1,\ldots,z_n) =$$

$$g(z_1,\ldots,z_n)[\xi(f_1(z_1,\ldots,z_n),\ldots,f_n(z_1,\ldots,z_n)) - (1-h(z_1,\ldots,z_n))\xi(0,\ldots,0)],$$

$$\mid z_i \mid \leq 1, i = 1,\ldots,n,$$

which is identical to (5.8).

□

Below we solve the functional equation (5.7). We first derive some preliminary results from (5.7). Noting that $\mathrm{E}(\mathrm{e}^{-\zeta\mathbf{I}}) = \alpha(\zeta,0)$ for $\mathrm{Re}\,\zeta \geq 0$ and $\mathrm{E}(\mathrm{e}^{-\omega\mathbf{B}}) = \alpha(0,\omega)$ for $\mathrm{Re}\,\omega \geq 0$,

$$\mathrm{E}\mathbf{C} = \frac{s+\frac{\gamma(\lambda)}{\lambda}}{1-\rho}; \tag{5.9}$$

$$\mathrm{E}\mathbf{I} = \frac{\gamma(\lambda)}{\lambda}; \tag{5.10}$$

$$\mathbf{EB} = \frac{s + \rho\dfrac{\gamma(\lambda)}{\lambda}}{1-\rho}; \tag{5.11}$$

$$\mathbf{EB}^2 = \frac{s^{(2)} + (2\rho s + \sum_{i=1}^{n} \lambda_i \beta_i^{(2)})\mathbf{EC}}{1-\rho^2}. \tag{5.12}$$

Remark 5.3 We may also obtain (5.10) directly by observing

$$\mathrm{E}(\mathbf{I} \mid \mathbf{I} > 0) = \frac{1}{\lambda},$$

while

$$\Pr\{\mathbf{I} > 0\} = \Pr\{(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (0, \ldots, 0)\} = \int_{t=0}^{\infty} \mathrm{e}^{-\lambda t} d\Pr\{\mathbf{B} < t\} = \gamma(\lambda).$$

We may also obtain (5.9) directly from (3.1) and (5.10) by observing that $\pi = \dfrac{\mathbf{EI}}{\mathbf{EC}}$. □

We now solve the functional equation (5.7). Obviously it suffices to find an expression for $\gamma(\omega)$ for $\mathrm{Re}\,\omega \geq 0$, as substituting such an expression into (5.7) yields an expression for $\alpha(\zeta, \omega)$. Define $\delta(\omega) := \sum_{i=1}^{n} \lambda_i (1 - \beta_i(\omega))$ for $\mathrm{Re}\,\omega \geq 0$. Putting $\zeta = 0$ in (5.7),

$$\gamma(\omega) = \sigma(\omega)\left[\gamma(\delta(\omega)) - \frac{\gamma(\lambda)}{\lambda}\delta(\omega)\right], \qquad \mathrm{Re}\,\omega \geq 0. \tag{5.13}$$

Define recursively

$$\begin{aligned} \delta^{(0)}(\omega) &= \omega, && \mathrm{Re}\,\omega \geq 0; \\ \delta^{(k)}(\omega) &= \delta(\delta^{(k-1)}(\omega)), && \mathrm{Re}\,\omega \geq 0, k = 1, 2, \ldots . \end{aligned}$$

Iterating (5.13),

$$\gamma(\omega) = \prod_{k=0}^{M} \sigma(\delta^{(k)}(\omega))\gamma(\delta^{(M+1)}(\omega)) - \frac{\gamma(\lambda)}{\lambda} \sum_{k=0}^{M} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega)), \tag{5.14}$$

$$\mathrm{Re}\,\omega \geq 0, M = 1, 2, \ldots .$$

Lemma 5.1

i. $\lim_{M\to\infty} \delta^{(M)}(\omega) = 0$ for all ω with $\mathrm{Re}\,\omega \geq 0$.

ii. $\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega))$ converges for all ω with $\mathrm{Re}\,\omega \geq 0$.

iii. $\sum_{k=0}^{\infty} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega))$ converges for all ω with $\mathrm{Re}\,\omega \geq 0$.

Proof
See appendix A. □

Lemma 5.1 implies, letting $M \to \infty$ in (5.14),

$$\gamma(\omega) = \prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega)) - \frac{\gamma(\lambda)}{\lambda} \sum_{k=0}^{\infty} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega)), \qquad \mathrm{Re}\,\omega \geq 0. \tag{5.15}$$

Putting $\omega = \lambda$ in (5.15),

$$\gamma(\lambda) = \frac{\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\lambda))}{1 + \frac{1}{\lambda} \sum_{k=0}^{\infty} \delta^{(k+1)}(\lambda) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\lambda))}. \tag{5.16}$$

Substituting (5.16) back into (5.15),

$$\gamma(\omega) = \prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega)) - \frac{\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\lambda))}{\lambda + \sum_{k=0}^{\infty} \delta^{(k+1)}(\lambda) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\lambda))} \sum_{k=0}^{\infty} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega)), \tag{5.17}$$

$$\mathrm{Re}\,\omega \geq 0.$$

Substituting (5.17) into (5.7) yields an expression for $\alpha(\zeta,\omega)$.
Finally some words on the joint past and residual lifetime distribution of a restricted cycle. This distribution will turn out to play a crucial role in the analysis of the waiting time distribution in section 6. Denote by $\mathbf{B}_P$ and $\mathbf{B}_R$ stochastic variables with the distribution of, respectively, the past and residual lifetime of a restricted cycle. From Cohen [10] p. 113 we have,

$$\mathrm{E}(\mathrm{e}^{-\omega_P \mathbf{B}_P - \omega_R \mathbf{B}_R}) = \frac{1}{\mathrm{E}\mathbf{B}} \frac{\gamma(\omega_R) - \gamma(\omega_P)}{\omega_P - \omega_R}, \qquad \mathrm{Re}\,\omega_P \geq 0, \mathrm{Re}\,\omega_R \geq 0. \tag{5.18}$$

In particular,

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{B}_P}) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{B}_R}) = \frac{1 - \gamma(\omega)}{\omega \mathrm{E}\mathbf{B}}, \qquad \mathrm{Re}\,\omega \geq 0. \tag{5.19}$$

From (5.19),

$$\mathrm{E}\mathbf{B}_P = \mathrm{E}\mathbf{B}_R = \frac{\mathrm{E}\mathbf{B}^2}{2\mathrm{E}\mathbf{B}}. \tag{5.20}$$

6 THE WAITING TIME IN THE GLOBALLY GATED POLLING SYSTEM

In section 5 in formula (5.18) we expressed the LST of the joint past and residual lifetime distribution of a restricted cycle into the LST of the restricted cycle time distribution. In the present section we relate the waiting time distribution to the joint past and residual lifetime distribution of a restricted cycle. Thus we obtain an expression for the LST of the waiting time distribution in terms of the LST of the restricted cycle time distribution.

We first introduce some notation. Denote by $\mathbf{V}_i(\mathbf{N})$ the total service time of $\mathbf{N}$ type-i customers, $i = 1,\ldots,n$, for any non-negative integer valued stochastic variable $\mathbf{N}$. So $\mathrm{E}(\mathrm{e}^{-\omega \mathbf{V}_i(\mathbf{N})}) = \mathrm{E}(\beta_i(\omega)^{\mathbf{N}})$, $\mathrm{Re}\,\omega \geq 0$, $i = 1,\ldots,n$. Denote by $\mathbf{A}_i(\mathbf{T})$ the number of type-i customers arriving during a period of length $\mathbf{T}$, $i = 1,\ldots,n$, for any non-negative real valued stochastic variable $\mathbf{T}$. So $\mathrm{E}(z_i^{\mathbf{A}_i(\mathbf{T})}) = \mathrm{E}(\mathrm{e}^{-\lambda_i(1-z_i)\mathbf{T}})$, $\mid z_i \mid \leq 1, i = 1,\ldots,n$. Denote by $\mathbf{B}_i$ and $\mathbf{S}_i$ stochastic variables having distribution, respectively, $B_i(\cdot)$ and $S_i(\cdot)$, $i = 1,\ldots,n$.

We now analyze the waiting time distribution of an arbitrary type-i customer, by distinguishing whether the customer arrives during a restricted cycle or during an idling period (thus terminating the idling period immediately by initiating a new restricted cycle), in other words, whether the customer sees the server working/switching or idling upon arrival. The waiting time of an arbitrary type-i customer that arrives during a restricted cycle, $\mathbf{W}_i^{(\mathrm{B})}$, is composed of

i. the residual lifetime of the restricted cycle in which it arrives;

ii. the total service time of all customers that arrive at $Q_1,\ldots,Q_{i-1}$ during the same restricted cycle;

iii. the total service time of all customers that arrive at Q_i during the past lifetime of the restricted cycle in which it arrives;

iv. the total switch-over time experienced by the server, moving from Q_1 to Q_i;

i.e.,

$$\mathbf{W}_i^{(\mathrm{B})} \stackrel{d}{=} \mathbf{B}_R + \sum_{j=1}^{i-1} \mathbf{V}_j(\mathbf{A}_j(\mathbf{B}_P + \mathbf{B}_R)) + \mathbf{V}_i(\mathbf{A}_i(\mathbf{B}_P)) + \sum_{j=1}^{i-1} \mathbf{S}_j, \qquad i = 1,\ldots,n. \tag{6.1}$$

So, using (5.18),

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_i^{(\mathrm{B})}}) = \prod_{j=1}^{i-1} \sigma_j(\omega) \times \tag{6.2}$$

$$\int_{t_P=0}^{\infty} \int_{t_R=0}^{\infty} \mathrm{e}^{-\omega t_R} \prod_{j=1}^{i-1} \left\{ \mathrm{e}^{-\lambda_j(1-\beta_j(\omega))(t_P+t_R)} \right\} \mathrm{e}^{-\lambda_i(1-\beta_i(\omega))t_P} d_{t_P,t_R} \Pr\{\mathbf{B}_P < t_P, \mathbf{B}_R < t_R\} =$$

$$\prod_{j=1}^{i-1} \sigma_j(\omega) \frac{1}{\mathbf{EB}} \frac{\gamma(\sum_{j=1}^{i} \lambda_j(1-\beta_j(\omega))) - \gamma(\sum_{j=1}^{i-1} \lambda_j(1-\beta_j(\omega)) + \omega)}{\omega - \lambda_i(1-\beta_i(\omega))}, \qquad i = 1,\ldots,n, \mathrm{Re}\,\omega \geq 0.$$

The waiting time of an arbitrary type-i customer that arrives during an idling period, $\mathbf{W}_i^{(\mathrm{I})}$, is composed solely of the total switch-over time experienced by the server, moving from Q_1 to Q_i, i.e.,

$$\mathbf{W}_i^{(\mathrm{I})} \stackrel{d}{=} \sum_{j=1}^{i-1} \mathbf{S}_j, \qquad i = 1, \ldots, n. \tag{6.3}$$

So,

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_i^{(\mathrm{I})}}) = \prod_{j=1}^{i-1} \sigma_j(\omega), \qquad i = 1, \ldots, n, \mathrm{Re}\,\omega \geq 0. \tag{6.4}$$

Combining (6.2) and (6.4), noting that an arbitrary customer, irrespective of which type, arrives during a restricted cycle and an idling period with probability $\mathbf{EB/EC}$ and $\mathbf{EI/EC}$ respectively,

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_i}) =$$

$$\prod_{j=1}^{i-1} \sigma_j(\omega) \frac{1}{\mathbf{EC}} \left[\mathbf{EI} + \frac{\gamma(\sum_{j=1}^{i} \lambda_j(1-\beta_j(\omega))) - \gamma(\sum_{j=1}^{i-1} \lambda_j(1-\beta_j(\omega)) + \omega)}{\omega - \lambda_i(1-\beta_i(\omega))} \right], \tag{6.5}$$

$$i = 1, \ldots, n, \mathrm{Re}\,\omega \geq 0.$$

Remark 6.1 For $n = 1$, using (5.9), (5.10) and (5.13), (6.5) reduces to

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}}) = \frac{(1-\rho)\omega}{\omega - \lambda(1-\beta(\omega))} \left[\frac{\mathbf{EI}}{s + \mathbf{EI}} + \frac{s}{s + \mathbf{EI}} \frac{1-\sigma(\omega)}{s\omega} \frac{\gamma(\omega)}{\sigma(\omega)} \right], \qquad \mathrm{Re}\,\omega \geq 0,$$

in which we recognize the well-known waiting time decomposition property of $M/G/1$ vacation models.

□

From (6.5), using (5.9) and (5.12),

$$\begin{aligned} \mathbf{EW}_i &= \left[1 + 2\sum_{j=1}^{i-1} \rho_j + \rho_i \right] \frac{\mathbf{EB}^2}{2\mathbf{EC}} + \sum_{j=1}^{i-1} s_j \\ &= \left[1 + 2\sum_{j=1}^{i-1} \rho_j + \rho_i \right] \frac{1}{1+\rho} \left[\frac{\sum_{j=1}^{n} \lambda_j \beta_j^{(2)}}{2(1-\rho)} + \frac{\rho s}{1-\rho} + \frac{s^{(2)}}{2(s + \frac{\gamma(\lambda)}{\lambda})} \right] + \sum_{j=1}^{i-1} s_j, \end{aligned} \tag{6.6}$$

$$i = 1, \ldots, n.$$

Remark 6.2 For $n = 1$ (6.6) reduces to

$$\begin{aligned} \mathbf{EW} &= [1+\rho]\frac{\mathbf{EB}^2}{2\mathbf{EC}} \\ &= \frac{\lambda\beta^{(2)}}{2(1-\rho)} + \frac{\rho s}{1-\rho} + \frac{s^{(2)}}{2(s+\frac{\gamma(\lambda)}{\lambda})}, \end{aligned}$$

which agrees with Takagi [19] p. 213 formula (5.40b).

□

Remark 6.3 Noting that $\sum_{i=1}^{n} \rho_i[1+2\sum_{j=1}^{i-1}\rho_j+\rho_i] = \rho(1+\rho)$, we may obtain from (6.6) the pseudo-conservation law for the model under consideration,

$$\begin{aligned} \sum_{i=1}^{n}\rho_i\mathbf{EW}_i &= \rho(1+\rho)\frac{\mathbf{EB}^2}{2\mathbf{EC}} + \sum_{i=1}^{n}\rho_i\sum_{j=1}^{i-1}s_j \\ &= \rho\frac{\sum_{j=1}^{n}\lambda_j\beta_j^{(2)}}{2(1-\rho)} + \frac{\rho^2 s}{1-\rho} + \rho\frac{s^{(2)}}{2(s+\frac{\gamma(\lambda)}{\lambda})} + \sum_{i=1}^{n}\rho_i\sum_{j=1}^{i-1}s_j. \end{aligned}$$

We may also obtain the pseudo-conservation law, without knowledge of the individual mean waiting times, using (5.9) and (5.10), from (3.20) with $GG = \{1,\ldots,n\}$, $\pi_1' = \dfrac{\mathbf{EI}}{\mathbf{EC}} = \dfrac{\gamma(\lambda)}{\lambda}\dfrac{1-\rho}{s+\frac{\gamma(\lambda)}{\lambda}}$, $\pi_1'' = 0$, $\pi_i = 0$, $i \neq 1$.

□

Remark 6.4 As seen from (6.6), the ordering of the queues involves an ordering of the mean waiting times,

$$\mathbf{EW}_{i+1} = \mathbf{EW}_i + [\rho_i+\rho_{i+1}]\frac{\mathbf{EB}^2}{2\mathbf{EC}} + s_i, \qquad i = 1,\ldots,n-1.$$

The ordering of the queues even involves a stochastic ordering of the waiting times, as seen from the derivation of (6.5).

On the one hand the ordering of the waiting times may be argued to be unfair. Elevator polling, recently introduced by Altman, Khamisy, & Yechiali [1], to some extent meets this objection to the globally gated service discipline. In elevator polling the server alternatingly passes through 'up' cycles, visiting the queues in the order $Q_1,\ldots,Q_n$, and 'down' cycles, visiting the queues in the order $Q_n,\ldots,Q_1$. Thus as an immediate advantage elevator polling saves the switch-over time from Q_n to Q_1. Under the globally gated service discipline, at the start of each cycle, up or down, a similar centralized gating procedure is executed as described before, so alternatingly Q_1 and Q_n function as home base. Similarly to [1], assuming the switch-over time from Q_{i+1} to Q_i to have the same distribution as the switch-over time from Q_i to Q_{i+1}, in elevator polling with a dormant server the mean waiting times at all queues may be shown to be equal, irrespective of the traffic characteristics. Obviously the waiting

time *distributions* at all queues are not equal. E.g., the variance of the waiting times is likely to be larger at the queues visited in the beginning or the end of a cycle than at the queues visited in the middle of a cycle.
On the other hand the ordering of the waiting times may be exploited to effectuate some kind of priorization. Following the line of this idea, similarly to Boxma, Levy, & Yechiali [7] simple index rules may be established for both static and dynamic optimization of the system performance, measured in terms of the mean waiting times.

□

As a justification of the dormant server policy, we now show the waiting time at each of the queues to be smaller (in the increasing convex ordering sense) than in the ordinary non-dormant server case. Let us label the stochastic variables and the associated LST's and pgf's in the dormant and non-dormant server case with a hat and a tilde respectively. From [7] we have

$$\begin{aligned} \mathrm{E}\tilde{\mathbf{W}}_i &= \left[1+2\sum_{j=1}^{i-1}\rho_j+\rho_i\right]\frac{\mathrm{E}\tilde{\mathbf{C}}^2}{2\mathrm{E}\tilde{\mathbf{C}}}+\sum_{j=1}^{i-1}s_j \\ &= \left[1+2\sum_{j=1}^{i-1}\rho_j+\rho_i\right]\frac{1}{1+\rho}\left[\frac{\sum_{j=1}^{n}\lambda_j\beta_j^{(2)}}{2(1-\rho)}+\frac{\rho s}{1-\rho}+\frac{s^{(2)}}{2s}\right]+\sum_{j=1}^{i-1}s_j, \qquad (6.7) \\ & \qquad\qquad i=1,\ldots,n. \end{aligned}$$

Subtracting (6.7) from (6.6),

$$\mathrm{E}\hat{\mathbf{W}}_i-\mathrm{E}\tilde{\mathbf{W}}_i=-\left[1+2\sum_{j=1}^{i-1}\rho_j+\rho_i\right]\frac{1}{1+\rho}\frac{\frac{\hat{\gamma}(\lambda)}{\lambda}}{s+\frac{\hat{\gamma}(\lambda)}{\lambda}}\frac{s^{(2)}}{2s}\leq 0, \qquad i=1,\ldots,n.$$

The waiting times are however not only smaller in mean, but also in the increasing convex ordering sense, i.e., $\mathrm{E}f(\hat{\mathbf{W}}_i)\leq\mathrm{E}f(\tilde{\mathbf{W}}_i)$ for any non-decreasing convex function $f(\cdot)$.

Lemma 6.1
$\hat{\mathbf{W}}_i \leq_{\mathrm{icx}} \tilde{\mathbf{W}}_i, \qquad i=1,\ldots,n,$

i.e., $\mathrm{E}f(\hat{\mathbf{W}}_i)\leq\mathrm{E}f(\tilde{\mathbf{W}}_i)$, $i=1,\ldots,n$, for any non-decreasing convex function $f(\cdot)$.

Proof
See appendix B.

□

The question remains whether the waiting times are also smaller in the stochastic sense, i.e., $\Pr\{\hat{\mathbf{W}}_i\leq t\}\geq\Pr\{\tilde{\mathbf{W}}_i\leq t\}$ for any $t\geq 0$.

7 THE QUEUE LENGTH IN THE GLOBALLY GATED POLLING SYSTEM

In section 5 we expressed the pgf of the joint queue length distribution at the beginning of a cycle into the LST of the restricted cycle time distribution. In the present section we relate the joint queue length distribution at polling epochs to the joint queue length distribution at the beginning of a cycle. Thus we obtain an expression for the pgf of the joint queue length distribution at polling epochs in terms of the LST of the restricted cycle time distribution. We study both the queue lengths at $Q_1, \ldots, Q_n$ at a polling epoch at Q_i, and the queue lengths at $Q_1, \ldots, Q_n$ seen by the server at successive polling epochs at $Q_1, \ldots, Q_n$.

First some words on the marginal queue length distribution. Denote by $\mathbf{L}_i$ the number of waiting customers at Q_i at an arbitrary epoch, $i = 1, \ldots, n$. Because of the PASTA property and an up- & down-crossing argument the number of customers at Q_i at an arbitrary epoch, $\mathbf{L}_i$, at a customer arrival epoch at Q_i, and at a customer departure epoch from Q_i are identically distributed, $i = 1, \ldots, n$, just as in an ordinary isolated $M/G/1$ queue. Thus $\mathrm{E}(z_i^{\mathbf{L}_i})$ may easily be found from $\mathrm{E}(z_i^{\mathbf{L}_i}) = \mathrm{E}(\mathrm{e}^{-\lambda_i(1-z_i)\mathbf{W}_i})\beta_i(\lambda_i(1-z_i)), \mid z_i \mid \leq 1, i = 1, \ldots, n$, where $\mathrm{E}(\mathrm{e}^{-\lambda_i(1-z_i)\mathbf{W}_i})$ is given by (6.5), again just as in an ordinary isolated $M/G/1$ queue, cf. Keilson & Servi [14].

In the remainder of the present section we analyze the joint queue length distribution at polling epochs. Denote by $\mathbf{Y}_{ij}$ the number of customers at Q_j at a polling epoch at Q_i, i.e., at the start of a visit to Q_i, $i = 1, \ldots, n, j = 1, \ldots, n$. Denote by $\mathbf{D}_i$ the indicator function of the event that an arbitrary customer (that arrives in an empty system) arrives at Q_i, i.e., a stochastic variable which is 1 with probability λ_i/λ and 0 with probability $1 - \lambda_i/\lambda$, $i = 1, \ldots, n$.

By the very nature of the globally gated service discipline $\mathbf{Y}_{ij}$ and $\mathbf{X}_j$, $i = 1, \ldots, n, j = 1, \ldots, n$, are related as follows.

$$\mathbf{Y}_{ij} \stackrel{d}{=} \mathbf{X}_j(i \leq j) + \mathbf{A}_j\Big(\sum_{k=1}^{i-1}(\mathbf{V}(\mathbf{X}_k) + \mathbf{S}_k)\Big) + \left(\mathbf{D}_j(i \leq j) + \mathbf{A}_j\Big(\sum_{k=1}^{i-1}\mathbf{D}_k\mathbf{B}_k\Big)\right)\mathrm{I}_{\{\mathbf{X}=0\}}$$

$$i = 1, \ldots, n, j = 1, \ldots, n, \qquad (7.1)$$

where $\mathrm{I}_{\{\mathbf{X}=0\}}$ denotes the indicator function of the event $(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (0, \ldots, 0)$ and $(i \leq j)$ is equal to 1 if $i \leq j$ and 0 if $i > j$. The notation was further introduced in section 6.

We first study the joint distribution of $\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in}$, the queue lengths at $Q_1, \ldots, Q_n$ at a polling epoch at Q_i, $i = 1, \ldots, n$. From (7.1),

$$\mathrm{E}(z_1^{\mathbf{Y}_{i1}} \ldots z_n^{\mathbf{Y}_{in}} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n) = \prod_{k=1}^{i-1}\sigma_k(\epsilon(z)) \times \qquad (7.2)$$

$$\left[\prod_{k=1}^{i-1}\beta_k(\epsilon(z))^{\mathbf{X}_k}\prod_{k=i}^{n}z_k^{\mathbf{X}_k} - \left(1 - \sum_{k=1}^{i-1}\frac{\lambda_k}{\lambda}\beta_k(\epsilon(z)) - \sum_{k=i}^{n}\frac{\lambda_k}{\lambda}z_k\right)(\mathbf{X} = 0)\right],$$

$$i = 1, \ldots, n, \mid z_k \mid \leq 1, k = 1, \ldots, n.$$

Unconditioning (7.2), using (5.6)

$$\mathrm{E}(z_1^{\mathbf{Y}_{i1}} \dots z_n^{\mathbf{Y}_{in}}) = \prod_{k=1}^{i-1} \sigma_k(\epsilon(z)) \times \tag{7.3}$$

$$\left[\gamma(\sum_{k=1}^{i-1} \lambda_k(1-\beta_k(\epsilon(z)) + \sum_{k=i}^{n} \lambda_k(1-z_k)) - \frac{\gamma(\lambda)}{\lambda}\left(\sum_{k=1}^{i-1} \lambda_k(1-\beta_k(\epsilon(z))) + \sum_{k=i}^{n} \lambda_k(1-z_k)\right)\right],$$

$$i = 1, \dots, n, \mid z_k \mid \leq 1, k = 1, \dots, n.$$

For $i = n+1$, using (5.6) and interpreting $\mathbf{Y}_{n+1j}$ as $\mathbf{X}_j$, $j = 1, \dots, n$, (7.3) reduces to (5.8). From (7.3),

$$\mathrm{Cov}(\mathbf{Y}_{il}, \mathbf{Y}_{im}) = \lambda_l \lambda_m \times \tag{7.4}$$

$$\left[\mathrm{Var}(\sum_{k=1}^{i-1} \mathbf{S}_k) + \mathbf{EC} \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)} + (\mathbf{EB}^2 - (\mathbf{EC})^2)(\sum_{k=1}^{i-1} \rho_k + (i \leq l))(\sum_{k=1}^{i-1} \rho_k + (i \leq m))\right],$$

$$i = 1, \dots, n, l = 1, \dots, n, m = 1, \dots, n.$$

For $i = n+1$, using (5.12) and interpreting $\mathbf{Y}_{n+1j}$ as $\mathbf{X}_j$, $j = 1, \dots, n$, (7.4) reduces to

$$\mathrm{Cov}(\mathbf{X}_l, \mathbf{X}_m) = \lambda_l \lambda_m (\mathbf{EB}^2 - (\mathbf{EB})^2), l = 1, \dots, n, m = 1, \dots, n,$$

which may also be obtained from (5.6).

We finally study the joint distribution of $\mathbf{Y}_{11}, \dots, \mathbf{Y}_{nn}$, the queue lengths at $Q_1, \dots, Q_n$ seen by the server at successive polling epochs at $Q_1, \dots, Q_n$. From (7.1),

$$\mathrm{E}(z_1^{\mathbf{Y}_{11}} \dots z_n^{\mathbf{Y}_{nn}} \mid \mathbf{X}_1, \dots, \mathbf{X}_n) = \prod_{i=1}^{n} \sigma_i(\sum_{k=i+1}^{n} \lambda_k(1-z_k)) \times \tag{7.5}$$

$$\left[\prod_{i=1}^{n} \beta_i(\sum_{k=i+1}^{n} \lambda_k(1-z_k))^{\mathbf{X}_i} \prod_{i=1}^{n} z_i^{\mathbf{X}_i} - \left(1 - \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} z_i \beta_i(\sum_{k=i+1}^{n} \lambda_k(1-z_k))\right)(\mathbf{X} = 0)\right],$$

$$\mid z_i \mid \leq 1, i = 1, \dots, n.$$

Unconditioning (7.5), using (5.6),

$$\mathrm{E}(z_1^{\mathbf{Y}_{11}}, \dots, z_n^{\mathbf{Y}_{nn}}) = \prod_{i=1}^{n} \sigma_i(\sum_{k=i+1}^{n} \lambda_k(1-z_k)) \times \tag{7.6}$$

$$\left[\gamma(\sum_{i=1}^{n} \lambda_i(1 - z_i \beta_i(\sum_{k=i+1}^{n} \lambda_k(1-z_k)))) - \frac{\gamma(\lambda)}{\lambda}\left(\sum_{i=1}^{n} \lambda_i(1 - z_i \beta_i(\sum_{k=i+1}^{n} \lambda_k(1-z_k)))\right)\right],$$

$$\mid z_i \mid \leq 1, i = 1, \dots, n.$$

From (7.6),

$$\mathrm{Cov}(\mathbf{Y}_{ll}, \mathbf{Y}_{mm}) = \lambda_l \lambda_m \times \tag{7.7}$$

$$\left[\mathrm{Cov}(\sum_{i=1}^{l-1} \mathbf{S}_i, \sum_{i=1}^{m-1} \mathbf{S}_i) + \mathbf{EC} \left(q(l,m) + \sum_{i=1}^{\min(l,m)-1} \lambda_i \beta_i^{(2)} \right) + (\mathbf{EB}^2 - (\mathbf{EC})^2)(\sum_{i=1}^{l-1} \rho_i + 1)(\sum_{i=1}^{m-1} \rho_i + 1) \right],$$

$$l = 1, \ldots, n, m = 1, \ldots, n,$$

where $q(l,m)$ is equal to β_l if $l < m$, β_m if $l > m$ and 0 if $l = m$.

Remark 7.1 Unlike in the ordinary non-dormant server case, cf. Boxma, Weststrate, & Yechiali [9], neither the queue lengths at a polling epoch, nor the queue lengths seen by the server at successive polling epochs need to be positively correlated. E.g., if all service times and switch-over times are zero, then (7.4) and (7.7) reduce to

$$\mathrm{Cov}(\mathbf{Y}_{il}, \mathbf{Y}_{im}) = -\frac{\lambda_l \lambda_m}{\lambda^2} (i \leq l)(i \leq m), \qquad i = 1, \ldots, n, l = 1, \ldots, n, m = 1, \ldots, n,$$

and

$$\mathrm{Cov}(\mathbf{Y}_{ll}, \mathbf{Y}_{mm}) = -\frac{\lambda_l \lambda_m}{\lambda^2}, \qquad l = 1, \ldots, n, m = 1, \ldots, n,$$

respectively. This may also be seen immediately, as during every cycle exactly one customer is served, which occurs at Q_i with probability λ_i/λ, $i = 1, \ldots, n$.

□

8 Conclusion

We studied a polling system with a dormant server, i.e., a polling system in which the server may be allowed to make a halt at a queue when there are no customers present in the system. In the first part of the paper we derived a pseudo-conservation law for a general model and used it to compare the dormant and the non-dormant server case. We further addressed the question at which queues the server should make a halt to minimize the mean total amount of work in the system. In the second part we directed the attention in particular to a globally gated polling system in which the server makes a halt at its home base when there are no customers present in the system. In the latter case the waiting time at each of the queues was shown to be smaller (in the increasing convex ordering sense) than in the ordinary non-dormant server case.

In the present paper we allowed the server only to make a halt at a queue when there are no customers present in the system. In fact we may allow the server also to make a halt at a queue in other cases when there are customers present in the system. A first option might be to maintain the service disciplines at the various queues, but to decide at the completion of *each* visit whether to switch or to idle, and not only at the completion of a visit that leaves the entire system empty. A second option might be also to drop the service disciplines at the various queues, and to decide at the completion of each *service* whether to serve another

customer if present, to switch, or to idle, like Liu, Nain, & Towsley [16].
Once having enlarged the freedom of decisions in the operation of the system, it is quite natural to consider the problem of finding a strategy that optimizes the performance of the system. As the enlarged freedom of decisions will complicate the analysis even further, there is little hope to solve the problem analytically. Remember that the pseudo-conservation law did not even hold enough information to solve the more specific problem at which queues the server should make a halt to minimize the mean total amount of work in the system. However, like the latter more specific problem, the problem of finding a strategy that minimizes $\sum_{i=1}^{n} c_i \lambda_i \mathrm{E}\mathbf{W}_i$ may still be handled as a semi-Markov decision problem.

Acknowledgement The author is grateful to O.J. Boxma for several valuable discussions, and to M. Eisenberg for giving him access to his notes containing an outline for determining the joint queue length distribution for the dormant server case with either the exhaustive or gated service discipline. Further the author is grateful to G.M. Koole and J.A.C. Resing for some useful suggestions concerning the proof of Lemma 6.1.

REFERENCES

[1] Altman, E., Khamisy A., Yechiali, U. (1992). On elevator polling with globally gated regime. *Queueing Systems* **11**, *Special Issue on Polling Models*, 85-90.

[2] Altman, E., Konstantopoulos, P., Liu, Z. (1992). Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems* **11**, *Special Issue on Polling Models*, 35-57.

[3] Blanc, J.P.C., Van der Mei, R.D. (1993). The power-series algorithm applied to polling systems with a dormant server. CentER Discussion Paper 9346. Tilburg University. Submitted to ITC-14.

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

[5] Boxma, O.J., Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic service systems. *J. Appl. Prob.* **24**, 949-964.

[6] Boxma, O.J., Groenendijk, W.P., Weststrate, J.A. (1990). A pseudo-conservation law for service systems with a polling table. *IEEE Trans. Commun.* **38**, 1865-1870.

[7] Boxma, O.J., Levy, H., Yechiali, U. (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Ann. Oper. Res.* **35**, 187-208.

[8] Boxma, O.J., Weststrate, J.A. (1989). Waiting times in polling systems with Markovian server routing. In: *Messung, Modellierung und Bewertung von Rechensystemen und Netzen*, eds. J. Stiege, J.S. Lie (Springer, Berlin), 89-104.

[9] Boxma, O.J., Weststrate, J.A., Yechiali, U. (1993). A globally gated polling system with server interruptions, and applications to the repairman problem. CWI Report BS-R9208. To appear in: *Probability in the Engineering and Informational Sciences.*

[10] Cohen, J.W. (1982). *The Single Server Queue* (North-Holland, Amsterdam, 2nd ed.).

[11] Eisenberg, M. (1971). Two queues with changeover times. *Oper. Res.* **19**, 386-401.

[12] Eisenberg, M. (1972). Queues with periodic service and changeover times. *Oper. Res.* **20**, 440-451.

[13] Groenendijk, W.P. (1989). Waiting-time approximations for cyclic service systems with mixed service strategies. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12*, ed. M. Bonatti (North-Holland, Amsterdam), 1434-1441.

[14] Keilson, J., Servi, L.D. (1990). The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Oper. Res. Letters* **9**, 239-247.

[15] Levy, H., Sidi, M. (1990). Polling systems: applications, modelling and optimization. *IEEE Trans. Commun.* **38**, 1750-1760.

[16] Liu, Z., Nain, P., Towsley, D. (1992). On optimal polling policies. *Queueing Systems* **11**, *Special Issue on Polling Models*, 59-83.

[17] Resing, J.A.C. (1993). Polling systems and multitype branching processes. CWI Report BS-R9128. To appear in: *Queueing Systems.*

[18] Takagi, H. (1986). *Analysis of Polling Systems* (The MIT Press, Cambridge (MA)).

[19] Takagi, H. (1991). *Queueing Analysis, Vol. 1* (North-Holland, Amsterdam).

[20] Takagi, H. (1990). Queueing analysis of polling models: an update. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 267-318.

[21] Takine, T., Hasegawa, T. (1990). On the $M/G/1$ queue with multiple vacations and gated service discipline. Preprint Department of Applied Mathematics and Physics, Kyoto University.

[22] Tijms, H.C. (1986). *Stochastic Modelling and Analysis: a Computational Approach* (Wiley, Chichester).

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

APPENDICES

A PROOF OF LEMMA 5.1

Lemma 5.1

i. $\lim_{M\to\infty} \delta^{(M)}(\omega) = 0$ for all ω with $\mathrm{Re}\,\omega \geq 0$.

ii. $\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega))$ converges for all ω with $\mathrm{Re}\,\omega \geq 0$.

iii. $\sum_{k=0}^{\infty} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega))$ converges for all ω with $\mathrm{Re}\,\omega \geq 0$.

Proof

For any stochastic variable $\mathbf{H}$, having distribution $H(\cdot)$ with LST $\eta(\cdot)$, holds

$$\begin{aligned} |\, 1-\eta(\omega)\,| &= |\int_{t=0}^{\infty} (1-\mathrm{e}^{-\omega t})dH(t)\,| \\ &\leq \int_{t=0}^{\infty} |\, 1-\mathrm{e}^{-\omega t}\,|\, dH(t) \\ &\leq \int_{t=0}^{\infty} |\,\omega t\,|\, dH(t) \\ &= \mathrm{E}\mathbf{H}\,|\,\omega\,|, \end{aligned} \tag{A.1}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$, since $|\, 1-\mathrm{e}^{-z}\,| \leq |\, z\,|$ for all z with $\mathrm{Re}\, z \geq 0$.

Proof of i. By induction,

$$\mathrm{Re}\,\delta^{(k)}(\omega) \geq 0, \qquad k = 0, 1, \ldots, \tag{A.2}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$.

Further, using (A.1),

$$\begin{aligned} |\,\delta(\omega)\,| &= |\sum_{i=1}^{n} \lambda_i(1-\beta_i(\omega))\,| \\ &\leq \sum_{i=1}^{n} \lambda_i\,|\,1-\beta_i(\omega)\,| \\ &\leq \sum_{i=1}^{n} \lambda_i\beta_i\,|\,\omega\,| \\ &= \rho\,|\,\omega\,|, \end{aligned}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$.

Hence, by induction, using (A.2),

$$|\,\delta^{(k)}(\omega)\,| \leq \rho^k\,|\,\omega\,|, \qquad k = 0, 1, \ldots, \tag{A.3}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$.

As $\rho < 1$ is assumed to hold, we conclude that $\lim_{M\to\infty} \delta^{(M)}(\omega) = 0$ for all ω with $\mathrm{Re}\,\omega \geq 0$.

Proof of ii. From the theory of infinite products, cf. Titchmarsh [23] p. 18, we have that $\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega))$ converges iff $\sum_{k=0}^{\infty} [1 - \sigma(\delta^{(k)}(\omega))]$ converges.
Using (A.1), (A.3), (A.2) and the assumption that $\rho < 1$,

$$\begin{aligned} |\sum_{k=0}^{\infty}[1-\sigma(\delta^{(k)}(\omega))]| &\leq \sum_{k=0}^{\infty} |1-\sigma(\delta^{(k)}(\omega))| \\ &\leq \sum_{k=0}^{\infty} s\,|\delta^{(k)}(\omega)| \\ &\leq s\sum_{k=0}^{\infty}\rho^k\,|\omega| \\ &= \frac{s}{1-\rho}\,|\omega| \\ &< \infty, \end{aligned}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$. So $\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega))$ converges for all ω with $\mathrm{Re}\,\omega \geq 0$.
Proof of iii. Because of (A.2),

$$|\sigma(\delta^{(k)}(\omega))| \leq 1, \qquad k = 0, 1, \ldots, \tag{A.4}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$. Using (A.3), (A.4) and the assumption that $\rho < 1$,

$$\begin{aligned} |\sum_{k=0}^{\infty}\delta^{(k+1)}(\omega)\prod_{l=0}^{k}\sigma(\delta^{(l)}(\omega))| &\leq \sum_{k=0}^{\infty}|\delta^{(k+1)}(\omega)|\prod_{l=0}^{k}|\sigma(\delta^{(l)}(\omega))| \\ &\leq \sum_{k=0}^{\infty}\rho^{k+1}\,|\omega| \\ &= \frac{\rho}{1-\rho}\,|\omega| \\ &< \infty, \end{aligned}$$

for all ω with $\mathrm{Re}\,\omega \geq 0$. So $\sum_{k=0}^{\infty} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega))$ converges for all ω with $\mathrm{Re}\,\omega \geq 0$.

□

B Proof of Lemma 6.1

Lemma 6.1
$\hat{\mathbf{W}}_i \leq_{\mathrm{icx}} \tilde{\mathbf{W}}_i, \qquad i = 1, \ldots, n,$

i.e., $\mathrm{E}f(\hat{\mathbf{W}}_i) \leq \mathrm{E}f(\tilde{\mathbf{W}}_i)$, $i = 1, \ldots, n$, for any increasing convex function $f(\cdot)$.

Proof
We sketch the line of the proof for the case $n = 1$, omitting the redundant indices. The proof for the case $n > 1$ follows from a straightforward generalization.

Consider Figure 1 and Figure 2; $\hat{\mathbf{V}}_t$ and $\tilde{\mathbf{V}}_t$ denote the amount of work in the system at time t in the dormant and non-dormant server case respectively. Stochastically the arrival, service, and switch-over processes in Figure 1 and Figure 2 are identical, but the realizations in both situations are *not* identical. However, the realizations are related as follows. In both situations the server incurs exactly the same switch-over times, although not in the same order. During one and the same switch-over time the arrival processes in both situations proceed synchronously. The same customer arrives at the same time (relative with regard to that switch-over time), requiring the same service time. Likewise the arrival processes in both situations proceed synchronously during one and the same service time. The Poisson nature of the arrival processes allows the realizations of the arrival, service, and switch-over processes in Figure 1 and Figure 2 to be typical. Consequently Figure 1 and Figure 2 depict typical realizations of the amount of work in the system at time t in the dormant and non-dormant server case respectively, as well as typical realizations of other induced stochastic processes, like waiting times and queue lengths.

At time $t = \mathbf{T}_0$ in both situations the system is empty and the server is at its home base Q, just back from switching. The server then starts switching for a time of length $\mathbf{S}_0$. $\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \ldots$ are independent stochastic variables with common distribution $S(\cdot)$. Customers arrive during the switch-over time $\mathbf{S}_0$ at (relative) time $t = \mathbf{A}_1, t = \mathbf{A}_1 + \mathbf{A}_2, \ldots, t = \mathbf{A}_1 + \ldots + \mathbf{A}_\mathbf{K}$, requiring service times of length $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_\mathbf{K}$. $\mathbf{A}_1, \mathbf{A}_2, \ldots$ are independent exponentially distributed stochastic variables with mean $1/\lambda$. $\mathbf{B}_1, \mathbf{B}_2, \ldots$ are independent stochastic variables with common distribution $B(\cdot)$.
In the situation of Figure 1 at time $t = \mathbf{T}_0 + \mathbf{A}_1$ the server interrupts switching, suspending the remainder $\mathbf{S}_0 - \mathbf{A}_1$ of the switch-over time $\mathbf{S}_0$, and starts serving the newly arrived customer, just as if the server would have remained idling at Q from time $t = \mathbf{T}_0$ on, awaiting the new customer to arrive.
At time $t = \mathbf{T}_1$ the system is empty again and the server is back again at its home base Q (these events occurring simultaneously for the first time). The server then starts switching, resuming the switch-over time $\mathbf{S}_0$.
At time $t = \mathbf{T}_1 + \mathbf{A}_2$ the server again interrupts switching, suspending the remainder $\mathbf{S}_0 - \mathbf{A}_1 - \mathbf{A}_2$ of the switch-over time $\mathbf{S}_0$, and starts serving the newly arrived customer, again just as if the server would have remained idling at Q from time $t = \mathbf{T}_1$ on, awaiting the new customer to arrive.
In the situation of Figure 2 at time $t = \mathbf{T}_0 + \mathbf{A}_1$ the server just continues switching, disregarding the newly arrived customer. At time $t = \mathbf{T}_0 + \mathbf{S}_0$ the server finishes switching and starts serving the newly arrived customers. Among the customers that arrive during $\mathbf{S}_0$, let's say the $\mathbf{K}_0$-th customer is the customer with the largest number of generations offspring. In the situation of Figure 2 the server incurs exactly the same switch-over times $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_{\mathbf{L}_\mathbf{K}}$ as in the situation of Figure 1, however in the order $\mathbf{S}_0, \mathbf{S}_{\mathbf{L}_{\mathbf{K}_0-1}+1}, \ldots, \mathbf{S}_{\mathbf{L}_{\mathbf{K}_0}}, \mathbf{S}_1, \ldots, \mathbf{S}_{\mathbf{L}_{\mathbf{K}_0-1}}, \mathbf{S}_{\mathbf{L}_{\mathbf{K}_0}+1}, \ldots, \mathbf{S}_{\mathbf{L}_\mathbf{K}}$.
This order guarantees that the system is empty after the switch-over time $\mathbf{S}_{\mathbf{L}_{\mathbf{K}_0}}$, which will facilitate comparing the waiting times later on. Still this order does not bias the waiting times, as $\mathbf{K}_0$ does not depend on the switch-over times $\mathbf{S}_1, \ldots, \mathbf{S}_{\mathbf{L}_\mathbf{K}}$.

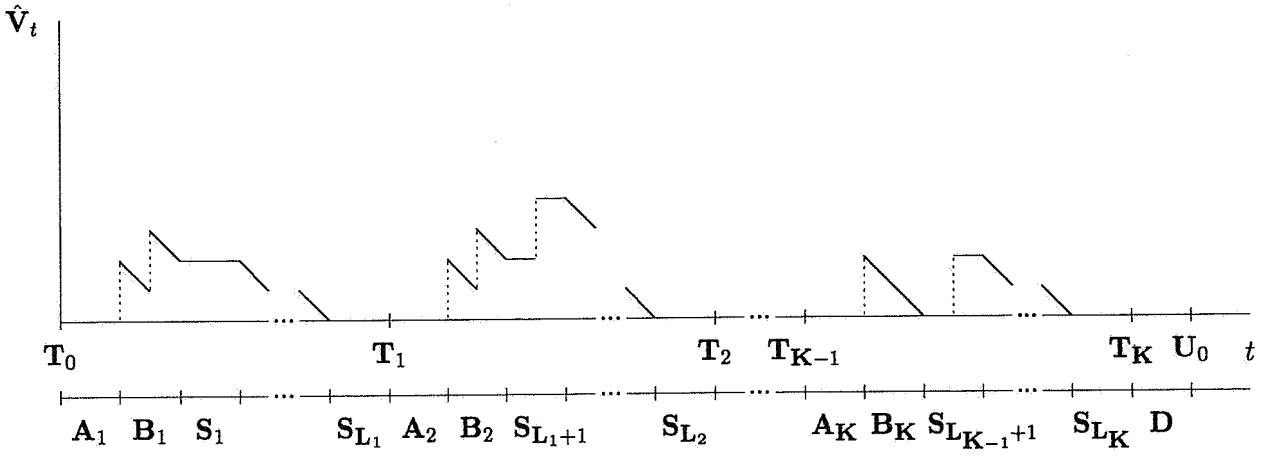

FIGURE 1. The amount of work in the dormant server case.

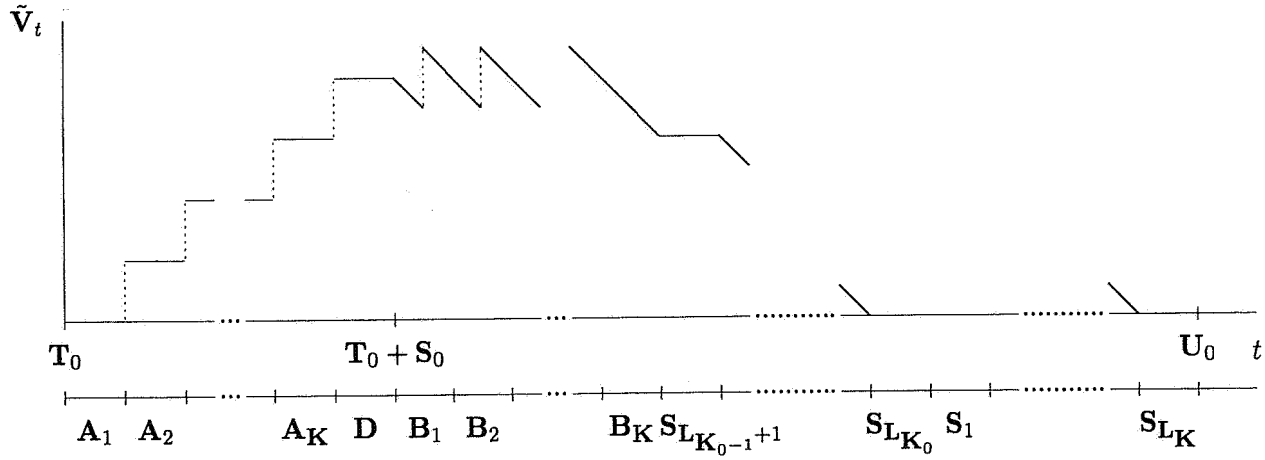

FIGURE 2. The amount of work in the non-dormant server case.

In the situation of Figure 1 at time $t = \mathbf{T_K}$ the system is empty and the server is back at its home base Q (these events occurring simultaneously for the $\mathbf{K}$-th time). The server then starts switching, resuming the switch-over time $\mathbf{S}_0$. At time $t = \mathbf{U}_0$ the server finishes switching, $\mathbf{U}_0 = \mathbf{T_K} + \mathbf{D}$, $\mathbf{D} = \mathbf{S}_0 - \mathbf{A}_1 - \ldots - \mathbf{A_K}$. Also in the situation of Figure 2 at time $t = \mathbf{U}_0$ the system is empty and the server just finishes switching. In the situation of Figure 2 the server has then incurred exactly the same switch-over times $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S_{L_K}}$ as in the situation of Figure 1, although not in the same order. The server has also served exactly the same customers, viz. the customers that arrive during $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S_{L_K}}$ and their offspring. Let's say their total number is $\mathbf{M}$. Concluding, at time $t = \mathbf{U}_0$ in both situations the system is empty and the server is at its home base Q, just back from switching.
Let $R^{(h)}$ be the h-th customer served from time $t = \mathbf{T}_0$ on in the dormant server case, $h = 1, 2, \ldots$. Denote by $\hat{\mathbf{W}}^{(h)}$ and $\tilde{\mathbf{W}}^{(h)}$ the waiting time of $R^{(h)}$ in the dormant and non-dormant server case respectively, $h = 1, 2, \ldots$. As the stochastic processes $\hat{\mathbf{W}}^{(h)}$, $h = 1, 2, \ldots$, and $\tilde{\mathbf{W}}^{(h)}$, $h = 1, 2, \ldots$, are regenerative with regard to $h = 1$ and $h = \mathbf{M} + 1$,

$$\mathrm{E}f(\hat{\mathbf{W}}) = \frac{1}{\mathrm{EM}}\mathrm{E}(\sum_{h=1}^{\mathbf{M}} f(\hat{\mathbf{W}}^{(h)})), \tag{B.1}$$

and

$$\mathrm{E}f(\tilde{\mathbf{W}}) = \frac{1}{\mathrm{EM}}\mathrm{E}(\sum_{h=1}^{\mathbf{M}} f(\tilde{\mathbf{W}}^{(h)})). \tag{B.2}$$

For comparing (B.1) and (B.2) we introduce some additional terminology. In the dormant server case the interval from time $t = \mathbf{T}_{k-1}$ to time $t = \mathbf{T}_k$ is referred to as the k-th busy interval, $k = 1, \ldots, \mathbf{K}$. Customers arriving during $\mathbf{S}_0$ (thus interrupting $\mathbf{S}_0$ in the dormant server case), together with their offspring, are called primary customers. The remaining customers, i.e., customers arriving during $\mathbf{S}_1, \ldots, \mathbf{S_{L_K}}$, together with their offspring, are called secondary customers.
With the busy intervals as background, the dormant and non-dormant server case differ in the service of primary customers, but not in the service of secondary customers. In the dormant server case the service of primary customers is balanced over the $\mathbf{K}$ busy intervals. In the non-dormant server case the service of primary customers is concentrated in the first busy interval, i.e., the counterpart of the $\mathbf{K}_0$-th busy interval in the dormant server case. Thus in the non-dormant server case the primary customers all bother one another and all bother the same secondary customers. This provides an intuitive feeling for the statement of the Lemma. A formal proof of the statement proceeds as follows.
Define $\mathbf{K}_1^{(lm)}$ ($\mathbf{K}_2^{(lm)}$) by $k \in \mathbf{K}_1^{(lm)}$ ($k \in \mathbf{K}_2^{(lm)}$) iff the k-th busy interval comprises at least l cycles and the l-th cycle comprises at least m services of primary (secondary) customers. Let $R_1^{(klm)}$ ($R_2^{(klm)}$) be the m-th primary (secondary) customer served in the l-th cycle in the k-th busy interval, $k \in \mathbf{K}_1^{(lm)}$ ($k \in \mathbf{K}_2^{(lm)}$). Denote by $\hat{\mathbf{W}}_1^{(klm)}$ and $\tilde{\mathbf{W}}_1^{(klm)}$ ($\hat{\mathbf{W}}_2^{(klm)}$ and $\tilde{\mathbf{W}}_2^{(klm)}$) the waiting time of $R_1^{(klm)}$ ($R_2^{(klm)}$) in the dormant and non-dormant server case respectively, $k \in \mathbf{K}_1^{(lm)}$ ($k \in \mathbf{K}_2^{(lm)}$).

Reordering summations,

$$\mathrm{E}(\sum_{h=1}^{\mathbf{M}} f(\hat{\mathbf{W}}^{(h)})) = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty}\left[\mathrm{E}(\sum_{k\in\mathbf{K}_1^{(lm)}} f(\hat{\mathbf{W}}_1^{(klm)})) + \mathrm{E}(\sum_{k\in\mathbf{K}_2^{(lm)}} f(\hat{\mathbf{W}}_2^{(klm)}))\right], \tag{B.3}$$

and

$$\mathrm{E}(\sum_{h=1}^{\mathbf{M}} f(\tilde{\mathbf{W}}^{(h)})) = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty}\left[\mathrm{E}\sum_{k\in\mathbf{K}_1^{(lm)}} f(\tilde{\mathbf{W}}_1^{(klm)})) + \mathrm{E}(\sum_{k\in\mathbf{K}_2^{(lm)}} f(\tilde{\mathbf{W}}_2^{(klm)}))\right]. \tag{B.4}$$

Define $\mathbf{K}^{(l)}$ by $k \in \mathbf{K}^{(l)}$ iff the k-th busy interval comprises at least l cycles. Denote by $\mathbf{C}^{(kl)}$ the length of the l-th cycle in the k-th busy interval, $k \in \mathbf{K}^{(l)}$.
For any waiting time $\mathbf{W}_p$ (cycle time $\mathbf{C}$), denote by $\mathbf{W}_{p1}$ ($\mathbf{C}_1$) the share constituted by the service times of primary customers, $p = 1, 2$. Denote by $\mathbf{W}_{p2}$ ($\mathbf{C}_2$) the remaining share in $\mathbf{W}_p$ ($\mathbf{C}$), i.e., the share constituted by the switch-over times $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_{\mathbf{L_K}}$ and the service times of secondary customers, $p = 1, 2$.
First we now compare the waiting times of the *primary* customers in the dormant and non-dormant server case. By construction $\hat{\mathbf{W}}_{11}^{(k11)} = 0$, $\tilde{\mathbf{W}}_{11}^{(k11)} = \sum_{h<k} \mathbf{C}_1^{(h1)}$, $\hat{\mathbf{W}}_{12}^{(k11)} = 0$, $\tilde{\mathbf{W}}_{12}^{(k11)} = \sum_{h>k} \mathbf{A}_h + \mathbf{D}$, for $k = 1, \ldots, \mathbf{K}$, and $\tilde{\mathbf{W}}_{11}^{(klm)} = \hat{\mathbf{W}}_{11}^{(klm)} + \sum_{\substack{h<k \\ h\in\mathbf{K}^{(l)}}} \mathbf{C}_1^{(hl)} + \sum_{\substack{h>k \\ h\in\mathbf{K}^{(l-1)}}} \mathbf{C}_1^{(hl-1)}$, $\hat{\mathbf{W}}_{12}^{(klm)} = \mathbf{C}_2^{(kl-1)}$, $\tilde{\mathbf{W}}_{12}^{(klm)} = \mathbf{C}_2^{(\mathbf{K}_0 l-1)}$, for $k \in \mathbf{K}_1^{(lm)}$, $l > 1$. (Note that $\mathbf{K}_1^{(11)} = \{1, \ldots, \mathbf{K}\}$, $\mathbf{K}_1^{(1m)} = \emptyset$, $m > 1$. Also note that $\mathbf{C}_2^{(\mathbf{K}_0 l-1)}$ exists if $k \in \mathbf{K}_1^{(lm)}$, $l > 1$, exists, as the $\mathbf{K}_0$-th customer among the customers arriving during $\mathbf{S}_0$ is the customer with the largest number of generations offspring.) As $\mathbf{C}_2^{(kl-1)} \stackrel{d}{=} \mathbf{C}_2^{(\mathbf{K}_0 l-1)}$, for $k \in \mathbf{K}_1^{(lm)}$, $l > 1$, in fact $\tilde{\mathbf{W}}_{12}^{(klm)} \stackrel{d}{=} \hat{\mathbf{W}}_{12}^{(klm)} \stackrel{d}{=}$ (say) $\mathbf{C}_2^{(l-1)}$, for $k \in \mathbf{K}_1^{(lm)}$, $l > 1$, with $\stackrel{d}{=}$ denoting equality in distribution. Using these observations,

$$\mathrm{E}(\sum_{k\in\mathbf{K}_1^{(lm)}} f(\hat{\mathbf{W}}_1^{(klm)})) = \mathrm{E}(\sum_{k\in\mathbf{K}_1^{(lm)}} f(\mathbf{C}_2^{(l-1)} + \hat{\mathbf{W}}_{11}^{(klm)})), \tag{B.5}$$

and

$$\mathrm{E}(\sum_{k\in\mathbf{K}_1^{(lm)}} f(\tilde{\mathbf{W}}_1^{(klm)})) = \mathrm{E}(\sum_{k\in\mathbf{K}_1^{(lm)}} f(\mathbf{C}_2^{(l-1)} + \hat{\mathbf{W}}_{11}^{(klm)} + \mathbf{V}^{(kl)})), \tag{B.6}$$

with $\mathbf{C}_2^{(0)} = 0$, $\mathbf{V}^{(k1)} = \sum_{h<k} \mathbf{B}_h + \sum_{h>k} \mathbf{A}_h + \mathbf{D}$, for $k = 1, \ldots, \mathbf{K}$, and $\mathbf{V}^{(kl)} = \sum_{\substack{h<k \\ h\in\mathbf{K}^{(l)}}} \mathbf{C}_1^{(hl)} + \sum_{\substack{h>k \\ h\in\mathbf{K}_{(l-1)}}} \mathbf{C}_1^{(hl-1)}$, for $k \in \mathbf{K}_1^{(lm)}$, $l > 1$. As $\mathbf{V}^{(kl)} \geq 0$, the expression in (B.5) is majorized by the expression in (B.6) for any increasing function $f(\cdot)$.

Second we now compare the waiting times of the *secondary* customers in the dormant and non-dormant server case. By construction $\hat{\mathbf{W}}_{21}^{(klm)} = \mathbf{C}_1^{(kl)}$, $\tilde{\mathbf{W}}_{21}^{(klm)} = 1_k \sum_{\mathbf{C}_1^{(hl)}>0} \mathbf{C}_1^{(hl)}$, $\tilde{\mathbf{W}}_{22}^{(klm)} = \hat{\mathbf{W}}_{22}^{(klm)} =$ (say) $\mathbf{W}_{22}^{(klm)}$, for $k \in \mathbf{K}_2^{(lm)}$, with $1_k = 1_{\{k=\mathbf{K}_0\}}$; $1_{\{k=\mathbf{K}_0\}} = 1$ if $k = \mathbf{K}_0$ and $1_{\{k=\mathbf{K}_0\}} = 0$ if $k \neq \mathbf{K}_0$. As the $\mathbf{K}_0$-th customer among the customers arriving during $\mathbf{S}_0$ is the customer with the largest number of generations offspring, we know that $\mathbf{C}_1^{(kl)} = 0$, for $k = 1, \ldots, \mathbf{K}$, if $\mathbf{C}_1^{(\mathbf{K}_0 l)} = 0$. Using these observations,

$$\mathrm{E}\Big(\sum_{k\in\mathbf{K}_2^{(lm)}} f(\hat{\mathbf{W}}_2^{(klm)})\Big) = \mathrm{E}\Big(\sum_{\substack{k\in\mathbf{K}_2^{(lm)}\\ \mathbf{C}_1^{(kl)}=0}} f(\mathbf{W}_{22}^{(klm)})\Big) + \mathrm{E}\Big(\sum_{\substack{k\in\mathbf{K}_2^{(lm)}\\ \mathbf{C}_1^{(kl)}>0}} f(\mathbf{W}_{22}^{(klm)} + \mathbf{C}_1^{(kl)})\Big), \tag{B.7}$$

and

$$\mathrm{E}\Big(\sum_{k\in\mathbf{K}_2^{(lm)}} \mathrm{E}f(\tilde{\mathbf{W}}_2^{(klm)})\Big) = \mathrm{E}\Big(\sum_{\substack{k\in\mathbf{K}_2^{(lm)}\\ \mathbf{C}_1^{(kl)}=0}} f(\mathbf{W}_{22}^{(klm)})\Big) + \mathrm{E}\Big(\sum_{\substack{k\in\mathbf{K}_2^{(lm)}\\ \mathbf{C}_1^{(kl)}>0}} f(\mathbf{W}_{22}^{(klm)} + 1_k \sum_{\mathbf{C}_1^{(hl)}>0} \mathbf{C}_1^{(hl)})\Big). \tag{B.8}$$

Note that $\Pr\{k \in \mathbf{K}_2^{(lm)}\}$ does not depend on k (= (say) $\alpha^{(lm)}$), given $\mathbf{C}_1^{(kl)} > 0$. Also note that the distribution of $\mathbf{W}_{22}^{(klm)}$, for $k \in \mathbf{K}_{22}^{(lm)}$, does not depend on k either ($\stackrel{d}{=}$ (say) $\mathbf{W}_{22}^{(lm)}$), given $\mathbf{C}_1^{(kl)} > 0$. So

$$\mathrm{E}\Big(\sum_{\substack{k\in\mathbf{K}_2^{(lm)}\\ \mathbf{C}_1^{(kl)}>0}} f(\mathbf{W}_{22}^{(klm)} + \mathbf{C}_1^{(kl)})\Big) = \alpha^{(lm)}\mathrm{E}\Big(\sum_{\mathbf{C}_1^{(kl)}>0} f(\mathbf{W}_{22}^{(lm)} + \mathbf{C}_1^{(kl)})\Big), \tag{B.9}$$

and

$$\mathrm{E}\Big(\sum_{\substack{k\in\mathbf{K}_2^{(lm)}\\ \mathbf{C}_1^{(kl)}>0}} f(\mathbf{W}_{22}^{(klm)} + 1_k \sum_{\mathbf{C}_1^{(hl)}>0} \mathbf{C}_1^{(hl)})\Big) = \alpha^{(lm)}\mathrm{E}\Big(\sum_{\mathbf{C}_1^{(kl)}>0} f(\mathbf{W}_{22}^{(lm)} + 1_k \sum_{\mathbf{C}_1^{(hl)}>0} \mathbf{C}_1^{(hl)})\Big). \tag{B.10}$$

For any convex function $f(\cdot)$, immediately from the definition, $\sum_{i\in I} f(x + x_i) \leq \sum_{i\in I} f(x + 1_{\{i=i_0\}} \sum_{h\in I} x_h)$, provided $i_0 \in I$ if $I \neq \emptyset$. As remarked before, we know that $\mathbf{C}_1^{(kl)} = 0$, for $k = 1, \ldots, \mathbf{K}$, if $\mathbf{C}_1^{(\mathbf{K}_0 l)} = 0$, i.e., $\mathbf{K}_0 \in \{k : \mathbf{C}_1^{(kl)} > 0\}$ if $\{k : \mathbf{C}_1^{(kl)} > 0\} \neq \emptyset$. So the expression in (B.9) is majorized by the expression in (B.10) for any convex function $f(\cdot)$.

□