Polling, production & priorities

DOI:
10.6100/IR628838

Document status and date:
Published: 01/01/2007

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Polling, Production
& Priorities

Thomas Stieltjes Institute
for Mathematics

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Winands, Erik M.M.

Polling, Production & Priorities / by E.M.M. Winands.
– Eindhoven : Technische Universiteit Eindhoven, 2007.
Proefschrift. – ISBN 978-90-386-1084-9
NUR 919
Subject headings : queueing theory, production systems
2000 Mathematics Subject Classification : 60K25, 68M20, 90B30

Printed by Ponsen & Looijen BV.
Cover by Paul Verspaget.

Polling, Production & Priorities

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op dinsdag 4 september om 16.00 uur

door

Erik Mathias Maria Winands

geboren te Valkenburg-Houthem

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. O.J. Boxma

en

prof.dr. A.G. de Kok

Copromotor:
dr.ir. I.J.B.F. Adan

Acknowledgements

I take great pleasure in acknowledging those individuals who contributed directly and indirectly to this document. First of all, I am indebted to my supervisors and (co-)promotors Ivo Adan, Onno Boxma, Geert-Jan van Houtum and Ton de Kok for their guidance, boundless enthusiasm and invaluable comments. Furthermore, I would like to thank Eitan Altman (INRIA), Alan Scheller-Wolf (Carnegie Mellon University), Mark Squillante (IBM) and Christian Timpe (BASF) for offering me the opportunity to perform part of the research at their organizations. I acknowledge the Netherlands Organisation for Scientific Research (NWO), that funded my stay at Carnegie Mellon University. Finally, thanks are due to all co-authors for giving permission to reuse the relevant parts of our joint work in the present monograph.

Erik Winands
June 2007

Contents

Chapter 1

Motivation

Imagine you are the production manager in a plant, where multiple standardized products are produced in a make-to-stock fashion. You are responsible for the production plan for the multi-stage production process. Within this process a single bottleneck machine dominates the scheduling decisions, which implies that you only have to draw up the production plan for this single machine as the production plans of upstream and downstream stages are easily deduced from this plan.

However, the development of such a production plan is certainly not as easy as falling off a log due to the fact that you are faced with the following complicating factors: finite production capacity, significant setup times and costs, a high utilization of capacity, and limited buffer capacity for end-products. Furthermore, demands, timing as well as amounts, for future periods are not known in advance, rather only statistical knowledge on these quantities is available. A final issue increasing the complexity of the planning process is the fact that production and setup times are variable due to breakdowns of the machine, human interference and the fact that raw material and tools are not always available.

As a production manager you probably get overwhelmed by the abundance of questions you come across: What product should I produce? When should I produce this product? In what quantity? Should I idle the machine? Should I keep stock for a product? And, if so, how much? How much buffer capacity should I dedicate to each product? And many, many more....

The example above is an illustration of the problem the present monograph focuses on: the so-called *stochastic economic lot scheduling problem* (SELSP). The SELSP deals with the make-to-stock production of multiple standardized products on a single machine with limited capacity under *random demands*, *possibly random setup times* and *possibly random production times*. The SELSP is a common problem in practice, e.g., in glass and paper production, injection molding, metal stamping and semi-continuous chemical processes, but also in bulk production of consumer products such as detergents and beers. Some specific applications described in the open literature are a laminate manufacturing plant (see Anupindi and Tayur [36]), a glass-containers manufacturing company (see Fransoo *et al.* [102]), a large consumer products manufacturer (see Gascon *et al.* [110]), a producer of plastic bumpers for cars (see Grasman *et al.* [114]) and an aerospace component supplier (see Sox and Muckstadt [196]). The author has been involved in several industry projects, among which two projects at a chemical company (see Winands *et al.* [P21] and Urlings [211]), a plant in an oleochemical production environment (see Ruigt [185]) and a producer of building materials (see Vught [215]). We touch upon these projects at various places in the present monograph.

Although the ubiquity of the SELSP reveals itself in a wide variety of applications, one can certainly observe common characteristics among these applications. The most pronounced similarity is the fact that a high utilization of capacity is typically prevalent due to either the magnitude of the setup times or the customer demands. A high utilization of capacity is oftentimes desired since the production facility is very expensive and produced products are commonalities, implying that a decrease in utilization would have a considerable effect on the cost price and would force the company out of business (see, also, Fransoo [101]).

In the SELSP, a production policy is needed which describes for each possible state of the system whether to continue production of the current product, whether to switch to another product or whether to idle the machine. The primary goal of such a production policy is the optimization of a pre-defined performance measure. Common performance measures in practice are, e.g., the average work-in-progress (WIP level), the minimization of total costs (sum of holding, backlogging and setup costs), the fraction of time that is lost due to setups, the average stock level or the waiting time of customers.

The development of such a policy for the SELSP is generally regarded as a challenging problem; the finite production capacity has to be *dynamically* distributed among the products in order to be reactive to the stochastic demands, processing and setup times (see Sox *et al.* [195]). The presence of *setup times* in combination with the *stochastic environment* are the key *complicating factors* of the problem. On the one hand, one aims for short cycle lengths, and thus frequent production opportunities for the various products, in order to be able to react to the stochasticity in the system. On the other hand, short cycle lengths increase the setup frequency, which has a negative influence on the amount of capacity available for production. Consequently, this effect hinders the timely fulfillment of demand.

The objective of the present monograph is the development and the analysis of mathematical models that capture the behavior and main effects of a number of production policies which are widely used in practical occurrences of the SELSP. Since the performance of a particular production schedule is, of course, highly contingent on the specific production environment in which it is deployed, we also pay attention to the identification of the characteristics of the environments which impact the performance. Finally, we propose and analyze extensions of the policies which lead to improved performance without losing the (organizational) advantages of the original policies.

The remainder of the present chapter is structured as follows. In Section 1.1 we give a detailed problem description of the SELSP after positioning the problem in the classical field of multi-product single-machine scheduling problems. Section 1.2 is devoted to the introduction of the class of policies on which the present study focuses, while the research statements of the present study are described in Section 1.3. Section 1.4 shows the contributions of the research in the present dissertation with respect to these statements.

1.1 Problem statement

The present section is divided into two parts. First, we position the SELSP in the framework of multi-product single-machine scheduling problems and, subsequently, we give a detailed problem description.

1.1.1 Multi-product single-machine scheduling problems

Scheduling production of multiple products on a single machine under tight capacity constraints is one of the classic problems in operations research. There are many variations of multi-product single-machine scheduling problems, but we may classify them by the following three characteristics:

1. *Presence or absence of setup times and/or costs.* The most important impact of

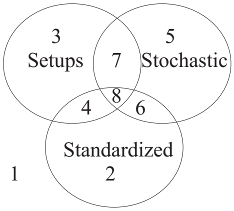


Figure 1.1: Multi-product single-machine scheduling problems.

setups on the production plan is that the products need to be produced in batches, since otherwise costly capacity is wasted on setups. Furthermore, setup times make it impossible to be completely responsive to the demand, as argued by Bourland [44].

2. *Customized or standardized products.* Since customized products can only be produced when there is a request for an order, these products have to be produced in a *make-to-order* fashion. In case of standardized products one may choose a *make-to-stock* production policy, because such products do not have to be produced to customer specifications. It is obvious that standardized products, thus, give more freedom in deciding when to make which product and in what quantity.

3. *Stochastic or deterministic environment.* In a completely deterministic environment one can confine oneself to a rigid production plan which is repeated over and over again. However, when the company has to be responsive to a stochastic environment such a rigid schedule will not suffice anymore.

The Venn diagram in Figure 1.1 depicts the eight subproblems, which are created by the above characterization. Of course, this diagram only provides a global classification of multi-product single-machine scheduling problems. One can easily think of examples of so-called hybrid systems (see, e.g., [P3], [27] and [94] for the combination of make-to-order and make-to-stock).

As introduced earlier, the present research considers the production of multiple standardized products on a single machine with limited capacity and setup costs under random demands, possibly random setup times and possibly random production times, depicted as subproblem 8 in Figure 1.1: the SELSP.

1.1.2 Problem description

We consider a system with a single production capacity for multiple products, in which there is unlimited stock space for each product and raw material is always available. Demands for the various products arrive according to stationary and mutually independent stochastic processes. We focus on single-item demand processes, but all results in the present monograph may be generalized to models with batch demand processes. Demand that cannot be satisfied directly from stock is backlogged until the product becomes available after production. The individual products are produced in a make-to-stock fashion with possibly stochastic production times. A possibly stochastic setup time occurs before the start of the production of a product. Motivated by the nowadays' efficient control of the production process, often the assumption is made that the production and setup times are deterministic. The setups are, furthermore, independent of the demand processes, production times and other setup times. Finally, only one product can be produced at a time.

Remark 1.1.1 In many practical situations products are produced on a multi-stage production line. If this line allows no interchangeability of products on other lines and if no intermediate storage is possible, such a line can be regarded as a single machine and, thus, the discussion of the present monograph fully applies (see Fransoo *et al.* [102]). □

1.2 The class of base-stock policies

In the past a plethora of policies for the SELSP has been developed and analyzed (see Chapter 2 for a classification and survey). The present section is devoted to the class of policies analyzed in the present study, the so-called *fixed-sequence base-stock policies.*

1.2.1 Fixed-sequence base-stock policies

In many firms encountering the SELSP, the following class of fixed-sequence base-stock policies is used for the control of the inventory of each product. We distinguish N products, which are numbered $1, 2, \ldots, N$. Subsequently, to each individual product a stock point is assigned which is controlled by a base-stock inventory policy. Under such a policy, for each product there exists a pre-defined desired number of items in stock, the base-stock level b_i, $i = 1, 2, \ldots, N$. When demand arrives at a stock point and the requested product is on stock, the demand is immediately fulfilled. Otherwise, demand is backlogged and fulfilled as soon as the product becomes available after production. A production order, also called replenishment order, is placed immediately after demand for the corresponding product has arrived. These production orders queue up at the production facility, where each product has its own designated queue.

On the strategy deployed by the production facility, the following two restrictions are imposed,

1. The products are produced according to a fixed production sequence;

2. When the machine starts production of a product, it will continue production until either the base-stock level has been reached or a second local criterion, i.e., only dependent on the stock level of the product currently setup, has been fulfilled.

Examples of local criteria are that only a limited number of items can be produced each production run or that the time for each production run is limited. Due to indistinguishability of the replenishment orders at the product facility (and the inability to measure the production time of an order before the start of production), one does not to have to worry about the scheduling discipline within the queue and, consequently, the *first come first served* (FCFS) discipline is assumed in the sequel. The combination of stock points and production facility is visualized in Figure 1.2.

The choice of the above class of policies is motivated by the following considerations. Firstly, this class of policies is widely used in practice in a wide variety of settings (see, e.g., Federgruen and Katalan [91; 93]). Secondly, the single production capacity that we consider may be a bottleneck in a larger multi-stage production process. Fixing the production sequence facilitates the coordination with downstream and upstream stages and leads to stability on the work floor, see also Bourland [44]. Thirdly, the fixed sequence is often imposed by sequence-dependent setup times implying that only a single sequence can be considered for implementation. Fourthly, base-stock policies are easily implementable and can effortlessly be monitored with minimal informational requirements, i.e., only *local* information of the product currently produced is needed (see, also, Chapter 2). Fifthly, fixed-sequence base-stock policies can be easily communicated to people on the work floor without having to resort to arcane terms. Sixthly, base-stock policies are natural generalizations of the cyclic policies advocated for deterministic cyclic problems (see, also, [91; 93]), which adds to the willingness of production managers to actually implement them on

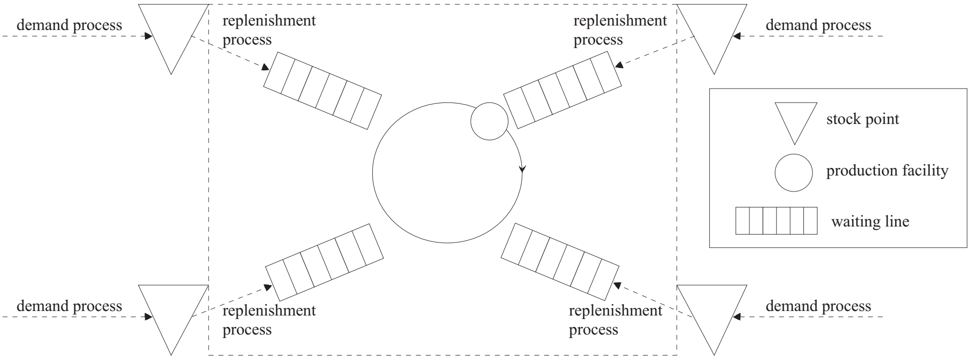


Figure 1.2: The fixed-sequence base-stock system.

the work floor. Finally, buffer capacities for end-products are often limited and since the net stock level cannot exceed its base-stock level, such a buffering constraint is satisfied by definition.

Example 1.2.1 A classical example of sequence-dependent setup times is a machine producing paint of all sorts of colors, which has to be cleaned between production runs of different colors. These cleaning times are highly dependent on the production sequence, e.g., producing a light color immediately after a dark one induces a huge cleaning time. In such settings, the production sequence is fixed to be the one minimizing total setup time in a cycle. □

1.2.2 Analysis of fixed-sequence base-stock policies

For given values of the base-stock levels b_i the steady-state net stock level N_i for product i is given by, for $i = 1, 2, \ldots, N$ (see Figure 1.3),

$$N_i = b_i - L_i, \tag{1.1}$$

where L_i denotes the steady-state *shortfall* (the number of outstanding production orders at the production facility) of product i. Notice that the net stock level of a product becomes negative, when the shortfall of this product is larger than its base-stock level.

One can verify that the shortfall of a product is independent of the base-stock levels implying that the performance of the production facility can be analyzed independently of these base-stock levels. Moreover, the shortfall distribution of a product at the production facility is identical to the queue length distribution of the corresponding queue in a so-called *polling system*. The arrival, service and setup time processes in such a polling system are identical to the demand, processing and setup time processes in the SELSP, respectively. For given base-stock levels, the evaluation of a fixed-sequence base-stock policy is, therefore, tantamount to evaluation of the corresponding polling system. Polling systems, which are the central queueing models of the present monograph, are formally defined below. For a picture of a polling system we refer to the area within the dotted box in Figure 1.2.

A typical polling system consists of a number of queues, attended by a single server in a fixed order. Customers arrive at all queues according to independent stochastic processes, while the service times at each queue are independent, identically distributed random variables. When the server starts service at a queue, a setup time is incurred. The successive setup times at a queue form identically distributed random variables. There

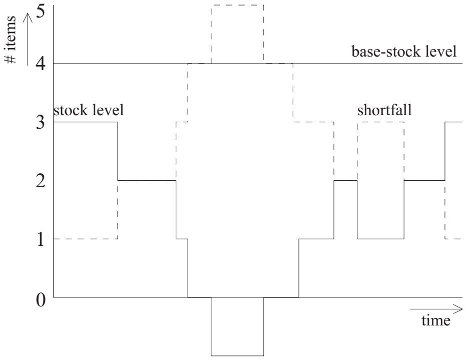


Figure 1.3: Example of the relation between base-stock level, net stock level and shortfall.

is a huge body of literature on polling systems that has continued to grow since the late 1950s, when the papers [155; 156] concerning a patrolling repairman model for the British cotton industry were published. Polling systems have a wide range of applications in communication, production, transportation and maintenance systems. Surveys on polling systems and their applications may be found in [203; 205; 206] and in [151]. The reader is referred to Chapter 2 for an overview of results on polling systems most relevant for the present study.

The analysis of polling systems can be done via various techniques, the main approaches being simulation and mathematical analysis. The present monograph makes mainly use of the latter because of the following considerations. Firstly, in the past people have observed an inefficiency of simulation techniques for polling systems, see, e.g., Blanc [41], which particularly reveals itself in the computation of queue length distributions as frequently needed throughout the present monograph. Secondly, in the SELSP a whole plethora of parameters influences system performance, whereas the production manager desires, nonetheless, to quantify the impact of each individual parameter separately. Mathematical analysis seems, therefore, the appropriate tool for evaluation and optimization, which is confirmed by the fact that many of the analytic results obtained in the present monograph explicitly reveal the impact of the individual input parameters on the performance measure of interest.

1.2.3 Construction of fixed-sequence base-stock policies

To introduce a fixed-sequence base-stock policy on the work floor, the production manager has to decide on the following three options:

1. How many items of each product should be produced per production run (*lot-sizing decision*)?

2. In which order and frequency should the products be produced (*sequencing decision*)?

3. What are the values of the optimal base-stock levels (*base-stock decision*)?

Below, these decisions are discussed in more detail.

Lot-sizing decision. This decision basically determines the length of the production runs and is, typically, made according to one of the following disciplines:

- *exhaustive* policy: when the machine starts production of a product, it will continue production until a pre-defined base-stock level has been reached;

- *gated* policy: when the machine starts production of a product, it will continue production until a production batch has been completed the size of which equals the difference between the base-stock level and the starting net stock level.

The main drawback of these traditional policies is the inability to prioritize among the different products for improving total system performance. More sophisticated lot-sizing policies offering this possibility, through bounding the lengths of the production runs, are

- *quantity-limited* policy: when the machine starts production of a product, it will continue production until either the base-stock level has been reached or a maximum number of items has been produced;

- *time-limited* policy: when the machine starts production of a product, it will continue production until either the base-stock level has been reached or a maximum amount of time has been spent on production.

In case of - practically relevant - deterministic processing times the quantity-limited and time-limited policy are identical to each other. In cases where the quantity and time limits equal infinity both policies are equivalent to the exhaustive service strategy.

Sequencing decision. This decision decides on the order and frequency in which products are produced. For example, if products A, B and C have to be produced a possible production sequence would be A-B-A-C. A sequence in which each product is produced exactly once in each cycle is called a *pure rotation sequence*, e.g., C-A-B. Unfortunately, in practice the production manager often has no rights to decide on the production sequence, since this is either imposed by the downstream and upstream stages or by strongly sequence-dependent setups.

Base-stock decision. The first two decisions determine the production strategy and, given these decisions, the shortfall distributions can be computed by analyzing the queue length distributions in the corresponding polling system. Since the control of the production facility is independent of the net stock level and the base-stock levels, optimization of the base-stock levels can be done separately from the analysis of (the shortfall in) the production facility. That is, under the assumptions of linear setup costs K_i and linear holding h_i and backlogging p_i costs, the total expected costs $Z(\cdot)$ may be written as a function of the vector of base-stock levels b, in the following way (for a pure rotation sequence)

$$Z(b) = \frac{\sum_{i=1}^{N} K_i}{\mathbb{E}[C]} + \mathbb{E}[\sum_{i=1}^{N} c_i\,(N_i)] = \frac{\sum_{i=1}^{N} K_i}{\mathbb{E}[C]} + \mathbb{E}[\sum_{i=1}^{N} c_i\,(b_i - L_i)], \tag{1.2}$$

with

$$c_i(x) = \begin{cases} h_i x, & x \geq 0, \\ -p_i x, & x < 0, \end{cases} \tag{1.3}$$

and where $\mathbb{E}[C]$ represents the mean cycle length which is dependent on the sequencing decisions, but which is independent of the base-stock and lot-sizing decisions. Now, the optimal base-stock levels can readily be obtained by solving standard newsboy problems using the computed queue length distributions (see Federgruen and Katalan [91; 92; 93]). That is, the optimal base-stock levels b_i^* are given by, for $i = 1, 2, \ldots, N$,

$$b_i^* = \min\{n \in \mathbb{N}_0 | P[L_i \leq n] \geq \frac{p_i}{p_i + h_i}\}. \tag{1.4}$$

For more information on newsboy problems, see, e.g., Zipkin [226]. In the remainder of the present monograph, it is assumed that the base-stock levels are optimized according to the above procedure.

The focus of the current research is mainly on the lot-sizing decision: What should the length of the production run be? Within the context of this lot-sizing decision the present monograph is, in particular, concerned with the evaluation and comparison of the traditional exhaustive and gated policies, on the one hand, and the more sophisticated quantity-limited policy, on the other hand. Evaluation and optimization of these lot-sizing disciplines are achieved through state-of-the-art analysis of several polling systems. It is important to remark that the lot-sizing and sequencing decisions made by the production manager in the implementation of the base-stock policy translate directly in the physical lay-out of these polling systems.

1.3 Research objectives

The present section discusses the two main research objectives studied. These objectives distinguish themselves in the lot-sizing policy analyzed, i.e., the exhaustive and gated policies, on the one hand, and the quantity-limited policies, on the other hand. Foregoing the survey of polling systems in Chapter 2, we want to highlight one of the most remarkable results in the polling literature to put our objectives in a proper light. That is, there exists a striking dichotomy in complexity between different polling systems, independently illuminated by Fuhrmann [104] and Resing [182]. If the lot-sizing discipline satisfies a certain branching property (as for the exhaustive and gated discipline), the polling system allows for an exact analysis by rather standard methods. Exact means that under the assumption of Poisson arrivals the generating functions of the queue length distributions can be expressed as infinite products, implying that even calculation of average queue lengths involve the solutions to sizeable systems of linear equations. If this branching property is, however, violated (as for the quantity-limited and time-limited lot-sizing policy), the corresponding polling systems can not be, or at least have not been, analyzed exactly in the general setting. Since the quantity-limited service policy is on the wrong side of the borderline, one has to resort to approximations for evaluation, let alone for optimization.

The first research objective deals with one of the most important performance measures, the exact evaluation of the average WIP level for each product (equalling the mean queue length in the corresponding polling system) under the most used lot-sizing policies, i.e., the exhaustive and gated strategies.

Research objective 1. *Development of a unifying exact framework for the analysis of the exhaustive and gated lot-sizing policies in terms of the average WIP levels under the assumption of Poisson demand processes.* □

Unifying in this context means that the framework should not only be able to compute the marginal average WIP levels, but should also be able to quantify the impact of the high utilization of capacity observed in practice (due to either the presence of setup times or the customer demands). It is shown that a major drawback of the exhaustive and gated lot-sizing policy is, however, that a single product, for which a high demand arrives in a certain period of time, may occupy the machine for quite a while. The impacts of this phenomenon on the other products are stock outs, highly variable cycle lengths and high costs. The quantity-limited lot-sizing policy circumvents this drawback and offers the possibility to the manager to control both the setup frequencies and the production runs (and, thus, the cycle lengths). One would expect that in this way the costs incurred at the machine can be significantly decreased. Startlingly, the quantity-limited policy has received no attention at all in the context of production-inventory systems, although this policy has won its spurs in the field of communication systems in the past (see, e.g., [43; 63]).

Evaluation of the quantity-limited lot-sizing policy is, thus, of great interest both for practitioners and theorists and, therefore, the following research objective is introduced.

Research objective 2. *Development of an efficient and accurate approximate tool for the analysis of the quantity-limited lot-sizing policy under the assumption of general demand processes.* □

We require that this approach, besides evaluating the performance of the quantity-limited policy, can also be applied to offer, albeit in an idealized mathematical setting, a preliminary exploration of an important managerial issue: *What is the gain in performance of bounding production runs by means of the quantity-limited lot-sizing policy in multi-item production-inventory systems?* Furthermore, it is important to stress that additional (qualitative) advantages of bounding the production runs are the facilitation of preventive maintenance and cleaning of the machine. Finally, the potential gains of the quantity-limited policy can be achieved without conceding to the (organizational) advantages of the class of base-stock policies as described in Section 1.2.

If cost benefits of the quantity-limited policy are supposedly high, such a tool is desired since then an accurate and efficient evaluation method contributes to the implementation of this policy on the work floor. In this respect, we recall that the exhaustive policy is a special case of the quantity-limited policy with an infinite quantity limit and may, therefore, ideally be evaluated within the same tool.

1.4 Contributions of the monograph

The present section describes our contributions with respect to the introduced research objectives. We present these contributions from the viewpoint of fixed-sequence base-stock policies. As argued before, the analysis of fixed-sequence base-stock policies is, essentially, tantamount to the analysis of polling systems. An overview of the existing literature on polling systems is given in Chapter 2, where we also relate our contributions to this field.

To put our contributions in a proper light it is important to remark, foregoing the survey on policies in the SELSP in Chapter 2, that in the context of fixed-sequence base-stock policies a series of papers by Federgruen and Katalan [91; 92; 93] has been published in the past. However, these papers deal with an *approximate* analysis of the exhaustive and gated lot-sizing policies and leave the quantity-limited policy untouched.

Ad research objective 1. In Chapter 3 an exact *Mean Value Analysis* (MVA) framework for the exhaustive and gated lot-sizing disciplines is presented, which computes the average WIP levels by exploiting direct mean value arguments. The merits of this MVA framework are its intrinsic simplicity and its intuitively appealing derivation. As a consequence, MVA may be applied, both in an exact and approximate manner, to a large variety of models. Within this framework the individual WIP levels can be efficiently obtained via the solution of a sparse set of linear equations, whereas for the total WIP level a closed-form expression is presented. The MVA framework allows the evaluation of both the exhaustive and gated discipline, but it is important to stress that the exhaustive policy is most relevant from a practical point of view, implying that throughout most quantitative and qualitative investigations are focussed on this discipline.

The MVA framework gives rise to explicit closed-form expressions, allowing for back-of-the-envelope calculations, for the individual WIP levels in the asymptotic regime of high utilization of capacity due to either customer demands or setup times. These expressions explicitly show the impact of all input parameters, yield insensitivity and monotonicity properties and unearth the (dis)similarities between the two sources of high utilization. In particular, it is shown that the exhaustive and gated lot-sizing disciplines display undesirable behavior if the utilization rate is high due to customer demand, which reveals itself, for example, in difficulties in the coordination between stages within the production process.

Motivated by the practical significance of the large setup times regime, we study this regime in more detail for a general class of branching-type lot-sizing policies by using more advanced techniques. The most remarkable result of this analysis is the fact that the stochastic system converges to its deterministic counterpart in the limit of increasing setup times implying that the sequencing decision is essentially tantamount to the one in the deterministic counterpart. The latter problem has been extensively studied in the seventies and eighties (see the survey of [86]). Moreover, a reasonable hypothesis, which may be formulated from this analysis, is that, in practice, production managers rely more on deterministic production strategies in production environments with significant setups than they do in environments with small setups in which stochastic (dynamic) policies seem to be more appropriate. Finally, the fact that the stochastic system becomes deterministic in the limit also implies that the exhaustive lot-sizing policy is optimal in terms of the WIP levels and that, thus, production runs should not be bounded in systems with extremely large setup times. For general settings, the latter conclusion does not always hold which we analytically show in the analysis of the second research objective. Finally, a challenging topic for further research related to our first research objective would be the classification of a general class of lot-sizing disciplines for which the MVA framework is applicable.

Ad research objective 2. In order to gain insights into the impact of bounding production runs and not to be diverted by other effects, Chapter 4 starts the analysis with a basic occurrence of the SELSP in an exact way. That is, we analyze a two-product system, in which a *high-priority* product is produced *exhaustively* and a low-priority product according to the *quantity-limited* service strategy. Ungainsayable, the two-queue model is, however, also of interest in its own right: Production applications, in which only two items have to be produced on a single production facility, are certainly not inconceivable. In such settings, it is quite natural to bound production runs to provide different service to the different items in order to improve system performance. In this model, we observe significant cost reductions by application of the quantity-limited policy, compared to the standard exhaustive policies, indicating the potential of the quantity-limited service discipline as lot-sizing rule in production environments. Foregoing the numerical evaluation in Chapter 4, we cite the example below which is one of the illustrative cases studied in this chapter.

Example 1.4.1 Consider a system with a single production capacity for two products, in which demands for the two products arrive according to Poisson processes with rate 0.375 and in which production and setup times for both products are exponentially distributed with means 1 and 0.25, respectively. Suppose that product 1 is a product with high costs, for which an exhaustive base-stock policy is implemented, whereas product 2 is of secondary importance compared to the first product and for which a quantity-limited base-stock policy is used. It turns out that the total costs in case the optimal quantity limit for product 2 is implemented are 35% lower than the total costs which would be incurred if a standard exhaustive policy were implemented for both products. □

The results obtained in the two-product case provide us with theoretical evidence that the quantity-limited strategy may lead to considerable cost reductions compared to the widely used (standard) exhaustive policy. Therefore, in Chapter 4 we develop an efficient and accurate approximate decomposition approach for the evaluation of quantity-limited lot-sizing policies under the most general imaginable assumptions, i.e., general number of products each with their own quantity limit in an environment with generally distributed arrival, service time and setup time distributions. The accuracy of the approximation scheme is verified by means of an extensive simulation study. The developed approximation approach turns out to be accurate, robust and computationally efficient. Therefore, it allows us to probe what happens should, for example, processing times change or demand for a product increase. Due to the low computational complexity of the developed procedure, it can be used in a large-scale study comparing the quantity-limited and exhaustive lot-sizing policies in the context of the SELSP. From an application point of view, it is important to remark that the algorithm can almost directly be applied to the lost sales case as well. Finally, since the approach is generic, it may serve as a basis for other types of queueing models that can be solved using the same technique.

The last part of Chapter 4 is devoted to a numerical simulation study assessing the quality of the quantity-limited lot-sizing policy as tool for prioritizing among products. It is shown that the quantity-limited lot-sizing policy outperforms the standard exhaustive policy leading to improvements in system performance for a variety of environments. In particular, the quantity-limited policy proves its worth in asymmetric production systems. Concluding we can say that the present monograph makes important methodological contributions in the evaluation of quantity-limited lot-sizing policies, optimization of such policies is however left for further research (although we present some explorative, yet preliminary, results in Chapter 4).

Finally, we should keep in the back of our mind that the results of the present monograph are certainly not limited to the described production setting, but may be used in the design and optimization phase of many other fields of application such as communication, maintenance, manufacturing and transportation as shown by the author, for example, in the papers [P4; P9].

We wish to end with an overview of the reports and papers upon which this thesis is largely built. Chapter 2 is partially based on the reports [P11; P12; P13; P21]. Chapter 3 stems from the journal papers [P5; P9; P10; P15], the conference papers [P4; P16] and the report [P17]. The basis of the material in Chapter 4 are the journal papers [P1; P8; P18], while some initial material was presented in the conference paper [P14]. The final chapter is based on the journal paper [P3; P6] and the reports [P2; P7].

Chapter 2

Literature review

The present chapter is divided into two parts. Firstly, we present a classification and survey on the policies studied so far for the SELSP (see Subsection 2.1). Secondly, we give an overview of the results on polling systems most relevant for the present monograph (see Subsection 2.2).

2.1 The stochastic economic lot scheduling problem

The aim of the present section is to give an overview of the research on the SELSP along with a comprehensive list of references. Therefore, we describe the most critical elements of a production plan, i.e., the sequencing and the lot-sizing policy. Based on these elements we propose a classification of the production strategies for the SELSP. The above two decisions are, of course, not the only decisions which ought to be made. The production manager has, for instance, to decide on the idle times between the production runs and the safety stocks as well. However, the sequencing and the lot-sizing decisions can be seen as the most critical ones.

The present section is a strongly abridged version of the literature review in the paper of Winands *et al.* [P13]. [P13] also contains, apart from a more extended literature survey, a comprehensive list of open problems in the SELSP, a thorough discussion of the relationship between the SELSP and its deterministic counterpart, suggestions for alternative policies not yet studied in the context of the SELSP and a brief overview of the recognition of the existence of the SELSP both in practice and in the academic society. For more details, the interested reader is referred to [P13].

Production sequence. The first critical element in a production plan is whether a *fixed production sequence* is used or not. A fixed production sequence means that there exists a pre-defined *order* and *frequency* for the production of the individual products. For production strategies using a fixed production sequence an additional classification can be made, i.e., whether or not a pre-defined *fixed cycle length* - which is the time between two successive completions of the production sequence - is used.

Thereupon, the production strategies for the SELSP can be divided into the following three categories:

- *Dynamic production sequence*;
- *Fixed production sequence in combination with a dynamic cycle length*;
- *Fixed production sequence in combination with a fixed cycle length.*

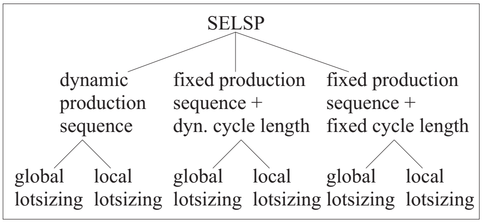

SELSP
dynamic production sequence
fixed production sequence + dyn. cycle length
fixed production sequence + fixed cycle length
global lotsizing
local lotsizing
global lotsizing
local lotsizing
global lotsizing
local lotsizing

Figure 2.1: Classification of the literature.

Lot-sizing policy. The second critical element of a production plan is the deployed lot-sizing policy, where we can distinguish two general classes of such policies:

1. *Global lot-sizing policies*: lot-sizing decisions may depend on the complete state of the system, i.e., stock levels of all the individual products and the state of the machine;

2. *Local lot-sizing policies*: lot-sizing decisions only depend on the stock level of the product currently setup.

Classification of the literature. By using the above two elements of a production plan a classification of the strategies for the SELSP can be made, which is depicted in Figure 2.1. Based on this classification, we review in the next subsection the existing literature on the SELSP. Surprisingly, until the end of the seventies the SELSP received almost no attention in the literature. Most likely, this lack of attention was not caused by an absence of practical interest, but by the intrinsic analytical complexity of the problem. To illustrate this feeling we cite here the following conclusion drawn by Vergin and Lee [213] in 1978 concerning the state of the research on the SELSP at that time:

> *The literature is almost completely void of not only the development of analytical models, but even of discussion of the problem. A thorough review of the production scheduling and inventory management journals and books would almost suggest that the scheduling problem does not exist. Yet the multiple product single machine system is quite common in industry and demand is inevitably stochastic.*

Due to this late start of the research on the SELSP, several interesting research questions are still unanswered as we see in the survey in the next subsection.

2.1.1 Strategy classes for the SELSP

2.1.1.A Dynamic production sequence. The papers concerning the two types of dynamic production sequence policies, i.e., using a global or a local lot-sizing policy, are described below.

Global lot-sizing policy. Sox and Muckstadt [196] propose a finite-horizon discrete-time stochastic optimization model under the assumption of the availability of overtime and deterministic production and setup times. They propose a method for finding optimal or near-optimal solutions applicable for small problems by using a Lagrangian decomposition algorithm. Sox and Muckstadt [196] assume that a setup for a product is incurred even if the same product was produced in the preceding period. They argue that this assumption can easily be relaxed at the expense of increased computation times.

Qiu and Loulou [179] present a multi-product model with limited stock space for each product under the assumption of backlogging, Poisson demand and deterministic production and setup times. They model the problem as a continuous-time semi-Markov decision problem with an infinite horizon, where the state space consists of the individual stock levels and the status of the machine. By using the successive approximations technique a policy can be derived on a truncated finite state space, which is then extended to a near-optimal policy for the original model. It is shown that a local lot-sizing policy is, in general, not optimal. Furthermore, Qiu and Loulou [179] conclude that for problems consisting of more than two products, their solution procedure cannot be both efficient and accurate due to the notorious curse of dimensionality. They suggest using composite products in these cases as in Graves [115]. Though they state that preliminary results indicate that this aggregation of products is promising, no results have appeared in the open literature yet.

Finally, we want to mention work of Karmarkar and Yoo [127], in which the SELSP is also studied under the assumption of backlogging, deterministic production and setup times. They present a formulation of a discrete-time stochastic dynamic programming problem over a finite horizon under the assumption of *time-varying* stochastic demand. Several Lagrangian relaxations of the problem are introduced, which can provide lower and upper bounds for the original problem. The results on small-scale problems are, however, not very encouraging.

Local lot-sizing policy. Production strategies that fall within this class are often of the so-called independent stochastic control type (see, e.g., Sox *et al.* [195]). This means that the lot-sizing decisions are made locally according to standard single product inventory control strategies such as (s, Q) or (s, S) policies, whereas the sequencing decisions are dynamically resolved by using priority rules. In particular, Zipkin [224] studies a continuous-time model under the assumption of backlogging, Poisson demand processes and generally distributed setup and production times. The individual lot-sizing policies are of the (s, Q) type, while the batches for the various products are produced in a *first come first served* (FCFS) order. Zipkin [224] makes the additional assumption that the production time of a batch is (nearly) independent of the size of this batch. He derives optimal batch sizes and reorder points with respect to total costs.

Winands *et al.* [P21] extend the analysis of [224] to general (renewal) arrival processes, multiple parallel lines and different service measures. Furthermore, [P21] does not make the assumption of the production times being independent of the batch sizes. [P21] develops a fast, accurate and easy-to-implement algorithm for the evaluation and optimization of the studied policy. The approach is implemented as a decision support tool in a chemical plant in Germany, which enabled them to make recommendations on the required inventory levels and tank capacities for the plant. By doing so, they identify major opportunities for improvement of current practice. Encouraged by this first positive application, the company intends to apply the tool inside and outside Germany. The case study is of particular interest for the present monograph since it shows a successful implementation of a queueing algorithm in practice, in all respects comparable to the tool developed in Chapter 4 for the quantity-limited lot-sizing policy, and illustrates how to take full advantage of the idiosyncrasies of a specific practical setting in such implementations.

Altiok and Shiue [29; 30; 31] study a model for the SELSP, in which an (s, S) policy is implemented for the lot-sizing decisions. The sequencing decision is made based on either one of the following two priority rules. The first one is a standard priority rule, which states that when the inventory position of the product currently setup reaches its base-stock level, the machine starts production of the highest priority product with inventory position below its reorder point. The second priority rule is a cyclical one, which serves the products with stock below their reorder points in a cyclical manner. In the case of backlogging, Altiok and Shiue [29] present an approximate analysis for the case of three products and in Altiok and Shiue [31] the extension is made to the N-product case. More specifically, an approximate iterative procedure is developed, which assumes independence among the stock levels of the products. In Altiok and Shiue [30], an exact analysis is performed for the lost sales case under the additional assumptions of phase-type distributed production times and exponentially distributed setup times.

Paternina-Arboleda and Das [178] analyze a base-stock policy, which works as follows: if the product currently being produced reaches its base-stock level, one is allowed to switch to another product or to keep the machine idle until a demand arrival epoch when switching may be worthwhile. For a small test bed, some improvements in system performance are reported compared to fixed-sequence base-stock policies. The methodology used in [178] can be summarized as a simulation optimization approach using reinforcement learning. A major drawback induced by this approach is that the resulting policy is very hard to implement; [178] (partially) solves this implementation problem by applying data mining classification techniques.

Brander *et al.* [55] present simulation results to assess the quality of deterministic lot-sizes in stochastic environments, which show that the determination of lot sizes is of less importance than, among other things, the sequencing decision. In [55] the latter decision is based on the run-out times of all products, where the run-out time of a product represents the expected time until this product runs out of stock.

2.1.1.B Fixed production sequence with a dynamic cycle length. In the context of fixed sequence strategies using a dynamic cycle length, we distinguish strategies using a global lot-sizing policy or a local one. These two types of strategies are discussed in more detail below.

Global lot-sizing policy. Markowitz *et al.* [157] propose a model, in which the demands are allowed to follow general renewal processes. At each point in time, the production manager has the following options: produce the product currently setup, idle the machine or switch to the next product in the production sequence. Markowitz *et al.* [157] study the cases of setup times and costs separately, but the combination can be analyzed without much additional effort. Motivated by well-known heavy-traffic limit theorems, a time-scale decomposition is made. By doing so, the SELSP can be approximated by a diffusion control problem. This problem can be solved explicitly for the setup cost case, whereas one has to resort to an algorithmic procedure in the setup time case. The paper is completed with a numerical evaluation of the resulting policies, in which, among others, a discrete-event simulation is used. In Markowitz and Wein [158] the same kind of heavy traffic analysis is applied to all kinds of related stochastic multi-product single-machine scheduling problems.

Bourland and Yano [45] use a two-level hierarchical policy for the SELSP under the assumption of deterministic production and setup times. Their strategy assumes that for each individual product a reorder point is given. Besides idle time and safety stocks, overtime may be used to respond to the stochastic demand. Since the backlogging costs are higher than the overtime costs, no demand is backlogged. At the upper level, the *planning* level, a cyclic schedule is obtained without neglecting the stochasticity of the demand. At this level one decides on the cycle length, stock levels and idle time allocations given the reorder points. Moreover, this planning level sets targets for the lower level, the *control* level. At this lower level, a control rule is defined that tries to follow the target schedule. The control policy does not alter the production sequence, but it may move the production starts forward or backward in time and, thus, the actual cycle length may differ from the target length. The production quantity is then determined by a so-called *match-up* lot-sizing policy, which is defined as follows in Bourland [44]: a match-up policy schedules production of a product in such a way that the stock level at the *planned* completion time

- and not necessarily at the *actual* completion time - of the production run will be equal to the base-stock level. In Bourland [44], it is shown by means of a simulation study that such a match-up policy follows the target cycles more effectively compared to a standard (exhaustive) base-stock policy. By using simulations Bourland and Yano [45] conclude that in comparison to safety stocks and overtime, idle times are a very expensive tool against demand uncertainty.

Gallego [107] proposes a three-level production strategy. At the first level, Gallego [107] constructs the production sequence, the production quantities and the idle times based on a deterministic procedure, but he does not add safety stocks yet. At the second level, Gallego [107] derives a policy which recovers the target schedule at minimal excess over average costs after a single disruption. A disruption may for instance be a machine failure, lack of raw materials, variations in demand or power shortages. The recovery of the target schedule is realized by adjusting the production quantities without altering the production sequence. The size of these adjustments for a product depends in general on the stock levels of all individual products. In Gallego [108] sufficient conditions are given for a base-stock recovery policy to be optimal. In case of a base-stock policy one only needs to monitor the stock level of the product currently setup. Several authors (see, e.g., Anupindi and Tayur [36] and Sox *et al.* [195]) see in these conditions a first step in the derivation of optimality conditions for the SELSP. At the third level, Gallego [107] adds safety stocks in order to efficiently use the control policy, that was shown to be optimal after a *single* disruption, in a stochastic environment.

Leachman and Gascon [143] develop a so-called *dynamic cycle lengths heuristic* in a discrete-time model under the assumption of non-stationary demand and deterministic production and setup times. The first step in their heuristic is the calculation of *target* cycle lengths in each review period via a deterministic approach by using moving averages of the demand forecasts. The second step is the determination of the *operational* cycle lengths, which are the minimal reductions of the target cycle lengths for which there is an adequate probability that these modified cycles can be maintained. This reduction is achieved by proportionally reducing the production quantities of all products in a cycle, while maintaining the fixed production sequence. The last step is the possible insertion of an idle period in a cycle, when all products have sufficient stock. Leachman *et al.* [144] improve the heuristics by increasing the lengths of the operational cycles, i.e., the operational cycles are closer to the target cycles, which results in lower costs and improved customer service. Furthermore, the decision rule concerning the insertion of an idle period is refined as well. In a later paper (Gascon *et al.* [110]), an extensive simulation study on the performance of the dynamic cycle lengths heuristic is undertaken. It is concluded that the performance of the heuristic is satisfactory as long as the load is not extremely high. Moreover, it turns out that the dynamic cycle length heuristic outperforms simple heuristics such as the EMQ rule.

Fransoo *et al.* [102] study a model for the SELSP under assumptions comparable to those of Leachman and Gascon [143] and Leachman *et al.* [144]. In particular, it is assumed that demand that cannot be fulfilled from stock is lost. Fransoo *et al.* [102] show numerically that the performance of the dynamic cycle lengths heuristic proposed by Leachman and Gascon [143] significantly decreases if the load increases. Therefore, an alternative heuristic is developed, that is able to keep the cycle lengths stable. By means of a simulation study, it is shown that this stable cycle length heuristic outperforms the dynamic cycle lengths heuristic when the total demand rate is close to or exceeds the production rate. Fransoo [100] and Fransoo *et al.* [102] give the following qualitative reason for this result. If one expects that a product will run out of stock in the forthcoming cycle, the dynamic cycle lengths heuristic will decrease the cycle length. Hence, the relative setup frequency increases and less capacity is available for production. Consequently, it becomes even more difficult to fulfill future demand.

Local lot-sizing policy. Anupindi and Tayur [36] present a fixed production sequence policy under the assumption of backlogging and deterministic production and setup times. They model a very general demand process, i.e., a non-Markovian compound demand process in which demand arrives for *sets* of products, and assume state-dependent setups. They allow a large number of lot-sizing strategies which are all variations of base-stock policies. A simulation-based procedure is developed to obtain the optimal base-stock levels for various performance measures. These performance measures contain not only the traditional *product-focussed* measures, such as costs and service levels based on individual products, but also *order-focussed* measures, like the order response time. Their paper is completed with numerical results for theoretical instances as well as for an industrial application. From these results, it clearly emerges that the popular product-focussed performance measures based on costs cannot be used as substitute for order-focussed measures.

Wagner and Smits [218] suggest a two-level continuous time model, where at the upper level an optimal fixed cycle schedule with respect to the expected setup and holding costs is derived. At the lower level a *periodic base-stock policy*, a so-called (R, S) policy, is used. The review periods are fixed and determined by the solution at the planning level, while the optimal base-stock levels are obtained by an algorithm developed by Smits *et al.* [194]. The planning and control levels are then instantaneously optimized by an integrative approach, which uses a local search optimization technique.

Vaughan [212] studies a model for the SELSP under the assumption of correlated demand, backlogging, deterministic production and setup times. His policy is characterized by an exhaustive base-stock policy and a target cycle length. This means that if a cycle is ended within the target length, the machine is idled. If not, the next cycle will commence immediately. Vaughan [212] concludes that demand correlation both increases the variance of the cycle length and causes correlation between the demand per period and the cycle length. Both effects lead to a higher variance of the total demand during a cycle and, thus, larger safety stock levels are needed compared to the uncorrelated demand case.

Recently, Eisenstein [85] introduced an extension of the fixed-sequence base-stock recovery policy of Gallego [107; 108]. The policy of [85] is more flexible than the policies of Gallego [107; 108] in that the policy of [85] is able to adjust the amount of idle time during recovery in response to disruptions. Numerical comparisons show that the policy of [85] is very effective compared to its competitors. Although this policy was designed for environments with a single disruption, it is numerically shown that it also is useful when disruptions are more pervasive (as also examined by [107; 108] for his policies).

Brander and Forsberg [53; 54] present an approximate method for the determination of safety stocks and base-stock levels for a given fixed production sequence, both with and without idle times. Moreover, they develop a control model making the decision whether to produce the next item in the sequence or to idle the machine. Although the lot-sizing discipline in [53; 54] is local, this control model incorporates global information. Finally, it is important to remark that [54] assumes that the production and setup times are deterministic, which is extended to stochastic production and setup times in [53].

The next strategy we describe is introduced by Federgruen and Katalan [91; 92; 93]. Besides the basic assumptions for the SELSP, the following additional assumptions are made in their model: unfulfilled demand is backlogged, the demand for a product follows a Poisson process and products are produced by an exhaustive or a gated base-stock policy. The production manager is allowed to insert a fixed idle time prior to the setup for a product in order to reduce the setup frequencies and, hence, the average setup costs. Federgruen and Katalan [91; 92; 93] show that the total average costs only depend on the *total* idle time inserted in a cycle and not on the *complete vector* of idle times. Hence, a strategy is completely specified by the vector of the base-stock levels and the total amount of idle time in a cycle. The optimal total idle time can be obtained by a numerical procedure. Finally, since the queue size distributions do not depend on the base-stock levels, these base-stock levels can be computed by solving standard newsboy problems. In Federgruen and Katalan

[93], their research is completed by the construction of an approximate optimal production sequence.

Grasman *et al.* [114] extend the exhaustive base-stock model of Federgruen and Katalan by adding random yields for the cases of backlogging, lost sales and expediting. In case of backlogging they derive similar newsboy equations in order to obtain optimal base-stock levels. In case of lost sales or expediting they have to resort to a heuristic for computing the (approximate) optimal base-stock levels. Krieg and Kuhn [134; 135] introduce continuous-time models for single-stage multi-product *Kanban systems*, which are completely identical to the SELSP with lost sales. It is assumed that demands arrive according to mutually independent Poisson processes and that the production and setup times are exponentially distributed. In Krieg and Kuhn [134], a system is analyzed with *state-independent* setups, whereas in Krieg and Kuhn [135] *state-dependent* setups are modeled, i.e., no setup for a product is incurred when there is no shortfall. Production quantities are in both models determined by an exhaustive base-stock policy. They decompose the multi-product Kanban system into multiple single-product single-server vacation models. The individual subsystems can be evaluated numerically by an approximate continuous-time Markov chain. Finally, it is noteworthy that the strategies of Federgruen and Katalan [91; 92; 93] and the above extensions [114; 134; 135] are some of the very few policies allowing for an analytical evaluation and optimization.

2.1.1.C Fixed production sequence with a fixed cycle length. Below we describe the two types of fixed production sequence strategies using a fixed cycle length.

Global lot-sizing policy. In the Master's thesis of Giezenaar [113], a case study of a chemical plant in the Netherlands is presented, for which a fixed production sequence strategy in combination with a fixed cycle length has been developed under the assumption of deterministic production and setup times. At the beginning of a cycle the production quantities are determined according to base-stock policies. When scheduling conflicts arise caused by the fixed cycle length, the production quantities are rationed in such a way that the production runs fit in this cycle length.

Local lot-sizing policy. Erkip *et al.* [87] introduce a discrete-time model under the assumption of backlogging, in which the production and setup times are deterministic. They propose a fixed cycle strategy, where fixed production times are allocated to products and where a base-stock policy is used. This means that not only the sequence and the total cycle length are fixed, but also the available capacity for each individual product is pre-defined. When the fixed amount of production time has expired, the product is not produced until the next cycle. Furthermore, when the product is on base-stock level before the end of the production interval, one does not switch to the next product in the sequence and, thus, the machine is idled. Their strategy is modeled as a quasi-birth-death process, which can be solved numerically by the matrix-analytic method.

Other recent work in this direction is by Bruin [58], who presents a generating function approach for the fixed cycle strategy under general traffic settings. Furthermore, a heavy-traffic approximation is developed for the optimal base-stock levels in case of Poisson demand distributions. Dellaert [78] mentions some drawbacks of such a fixed cycle strategy. The most important one is that no pooling effect is obtained by the fixed pre-allocation of production capacities to products. For a more detailed description of fixed cycle strategies in the context of make-to-order production situations, see Chapter 3 of Dellaert [78].

2.1.2 Position of the present monograph

Table 2.1 summarizes the classification of the literature discussed in Subsection 2.1.1. The fixed-sequence base-stock policies can be easily positioned in this classification, i.e.,

	Lot sizing strategy	
Production sequence	**Global**	**Local**
Dynamic	Karmarkar and Yoo [127] Qiu and Loulou [179] Sox and Muckstadt [196]	Altiok and Shiue [29], [30], [31] Brander *et al.* [55] Paternina-Arboleda and Das [178] Winands *et al.* [P21] Zipkin [224]
Fixed + dynamic cycle length	Bourland and Yano [44], [45] Fransoo *et al.* [102] Gallego [107], [108] Gascon *et al.* [110], [143], [144] Markowitz *et al.* [157], [158]	Anupindi and Tayur [36] Brander and Forsberg [53], [54] Eisenstein [85] Federgruen and Katalan [91], [92], [93] Grasman *et al.* [114] , [218] Krieg and Kuhn [134], [135] Smits *et al.* [194] Vaughan [212]
Fixed + fixed cycle length	Giezenaar [113]	Bruin [58] Erkip *et al.* [87]

Table 2.1: An overview of the classification.

this class of policies can be characterized by a fixed production sequence (with a dynamic cycle length) and a local lot-sizing policy. The contributions of the monograph to this area have been discussed, in detail, in Section 1.4.

2.2 Polling systems

The present monograph investigates several specific polling systems as modeling tools of fixed-sequence base-stock policies (see Chapter 1), which have some characteristics in common. To avoid duplication, the present section gives a general model description which is valid for the systems studied in subsequent chapters. For details which are specific for the systems considered, we refer to the corresponding chapters. Furthermore, we survey the state of the art in the analysis of polling systems, where it is certainly not our intention to present an encyclopedic overview of all available results. Instead, we want to illuminate the (mostly exact and asymptotic) concepts which put the contributions of the present monograph in the right perspective. For much more detailed surveys on polling systems and their applications, we refer to [203; 205; 206] and [151; 214].

First of all, to unify the presentation in the remainder of the present monograph all systems - unless specified otherwise - are described in terms of queues where customers arrive according to stochastic processes and receive service rather than products facing random demands which are produced to stock. Similarly, we adopt the nomenclature for the service policies as commonly used in the field of polling systems, e.g., we talk about the k-limited service policy rather than about the quantity-limited lot-sizing policy.

2.2.1 Basic model

We consider a system with a single server for $N \geq 1$ queues, in which there is infinite buffer capacity for each queue. The server visits and serves the queues in a fixed cyclic order. We index the queues by i, $i = 1, 2, \ldots, N$, in the order of the server movement. For compactness of presentation, all references to queue indices greater than N or less than 1 are implicitly assumed to be modulo N, e.g., queue $N+1$ actually refers to queue 1. Throughout the present monograph, it is assumed that within a queue customers are served FCFS. Obviously, the mean waiting times are the same under any work-conserving non-preemptive scheduling discipline (which excludes the creation and destruction of work see, e.g., [47]) that does not account for the actual service requests of the customers.

Customers arrive at all queues according to independent Poisson processes with rates λ_i, $i = 1, 2, \dots, N$, where the total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service times B_i at queue i are independent, identically distributed random variables with distribution function $F_i(\cdot)$, density $f_i(\cdot)$ and *Laplace Stieltjes Transform* (LST) $\beta_i(\cdot)$, $i = 1, 2, \dots, N$. The first two moments of the service time of an arbitrary customer are given by, respectively,

$$\mathbb{E}[B] = \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i]}{\Lambda}, \qquad \mathbb{E}[B^2] = \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i^2]}{\Lambda}. \tag{2.1}$$

When the server starts service at queue i, a setup time S_i is incurred with LST $\sigma_i(\cdot)$, $i = 1, 2, \dots, N$. The mean and the variance of the total setup time in a cycle are given by, respectively,

$$\mathbb{E}[S] = \sum_{i=1}^{N} \mathbb{E}[S_i], \qquad \mathbb{V}\text{ar}[S] = \sum_{i=1}^{N} (\mathbb{E}[S_i^2] - \mathbb{E}[S_i]^2). \tag{2.2}$$

These setup times are identically distributed random variables, independent of any other event involved. In particular, they are independent of the service times. Furthermore, it is assumed that a setup is incurred even if the subsequent queue is empty. For further reference, we introduce the mean residual service time and the mean residual setup time for queue i, which can be expressed as follows, respectively,

$$\mathbb{E}[R_{B_i}] = \frac{\mathbb{E}[B_i^2]}{2\mathbb{E}[B_i]}, \quad \mathbb{E}[R_{S_i}] = \frac{\mathbb{E}[S_i^2]}{2\mathbb{E}[S_i]}, \qquad i = 1, 2, \dots, N. \tag{2.3}$$

The occupation rate ρ_i (excluding setups) at queue i is defined by $\rho_i = \lambda_i \mathbb{E}[B_i]$ and the total occupation rate ρ is given by

$$\rho = \sum_{i=1}^{N} \rho_i. \tag{2.4}$$

The cycle length C_i of queue i, $i = 1, 2, \dots, N$, is defined as the time between two polling instants of this queue, where a polling instant of queue i is defined as the moment at which the server starts a visit at this queue (after a setup). It is well known that the mean cycle length is independent of the queue involved and is given by

$$\mathbb{E}[C] = \frac{\mathbb{E}[S]}{1-\rho}. \tag{2.5}$$

This identity can be proved by observing that the amount of work *arriving* during a cycle should on average equal the amount of work *departing* during a cycle, i.e.,

$$\rho \mathbb{E}[C] = \mathbb{E}[C] - \mathbb{E}[S]. \tag{2.6}$$

Unfortunately, higher moments of the cycle length are analytically intractable and, certainly, depend on the queue involved.

The visit period V_i of queue i, $i = 1, 2, \dots, N$, is the time the server spends servicing customers at queue i excluding setup time. Since the server is working a fraction ρ_i of the time on queue i, the mean of a visit period of queue i reads

$$\mathbb{E}[V_i] = \rho_i \mathbb{E}[C], \qquad i = 1, 2, \dots, N. \tag{2.7}$$

Subsequently, the intervisit period I_i of queue i, the time between a departure epoch of the server from queue i and its subsequent arrival to this queue, is defined as

$$I_i := C_i - V_i, \qquad i = 1, 2, \ldots, N. \tag{2.8}$$

Our main interest is in the waiting time W_i of a type-i customer, $i = 1, 2, \ldots, N$, which is defined as the time in steady state from a customer's arrival at queue i until the start of his service. Note that all results for the waiting time distribution can be readily translated into results for the queue length distribution - and vice versa - via the distributional form of Little's Law [130].

There exists a sharp and startling distinction in the complexity of the analysis of polling systems which has been independently illuminated by [104] and [182] via the use of a multi-type branching approach. That is, if the service discipline satisfies a certain branching property (defined below), the polling system allows for an exact analysis by rather standard methods. If this branching property is, however, violated, the corresponding polling system defies an exact analysis except for some special (two-queue and symmetric) cases. The branching property is defined as follows [104; 182],

Property 2.2.1 *If the server arrives at queue i to find l_i customers there, then during the course of the server's visit, each of these l_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function (PGF) $h_i(z) = h_i(z_1, \ldots, z_N)$, which can be any N-dimensional probability generating function.* □

The most important members of the class of policies satisfying Property 2.2.1 are

- *Exhaustive* policy: when the server polls a queue, he serves its customers until that queue is empty;
- *Gated* policy: when the server polls a queue, he serves all, and only, customers found at the polling instant.

Other policies for which Property 2.2.1 holds are the *binomial-exhaustive* [148] and the *binomial-gated* [149] disciplines.

Property 2.2.1 does not hold for the *k-limited* service strategy, where the server continues working at queue i, $i = 1, 2, \ldots, N$, until either a predefined number of k_i customers is served or until the queue becomes empty whichever occurs first. To see this, consider the simplest variant, the 1-limited discipline. At the end of the server's visit, the first served customer present at the polling instant has been effectively replaced by a population of all customers arrived during his service. The other customers present at a polling instant are not served at all and are each 'replaced' by a single customer at the queue under consideration. Consequently, not all customers are replaced in an i.i.d. manner and Property 2.2.1 is violated. Other policies violating Property 2.2.1 are, for example, the *Bernoulli* [129] and *time-limited* [43] disciplines. Finally, note that these exhaustive, gated and k-limited service disciplines correspond unambiguously to the exhaustive, gated and quantity-limited lot-sizing disciplines defined in the preceding chapter.

The remainder of the present section is divided into three parts. Firstly, we sketch how an exact analysis can be performed for the complete class of policies allowing a multi-type branching process interpretation via a classical generating function approach. Secondly, we survey the (limited) results for the k-limited policy, which can be seen as representative for the class of policies not being contained in the branching-type framework, and outline which unsurmountable difficulties occur in the extension to more general k-limited systems. Thirdly, we relate our contributions stated in Chapter 1 to the field of polling systems (both to systems with and without a multi-type branching structure).

In the present section, we outline all results for continuous-time cyclic systems with Poisson arrivals. However, we may extend all results, without seriously complicating the analysis, to discrete time, to periodic polling or to batch arrivals.

2.2.2 Branching-type policies

Throughout the present subsection, we assume that the service discipline at each queue satisfies Property 2.2.1 which implies that the joint queue lengths at polling instants can be represented by a multi-type branching process with immigration (see [104; 182]). In systems *with* setup times we are dealing with immigration *in each state*, whereas in systems *without* setup times immigration only takes place *in state zero* (when the whole system is empty). The theory of multi-type branching processes (see, e.g., [180]), which has been developed largely in the early seventies, is well-matured and provides us with necessary and sufficient ergodicity conditions and gives expressions for the generating function of the joint queue length process at polling instants. Building upon these results it turns out that one can derive pseudo-conservation laws, intensity-weighted sums of mean waiting times, and closed-form expressions for asymptotic performance measures in case of increasing load and/or increasing setup times.

Finally, it is important to remark that we allow different service disciplines at different queues and that we focus on *nonidling* service disciplines, i.e., the server never idles while at queue i if there is work in queue i, satisfying Property 2.2.1.

2.2.2.A Stability. The conditions $\rho < 1$ and $\mathbb{E}[S] < \infty$ constitute necessary and sufficient stability conditions for any nonidling policy that satisfies Property 2.2.1 with $h_i(z_1, \ldots, z_N) \neq z_i$ and, thus, $0 \leq \frac{\partial}{\partial z_i} h_i(z)|_{z=1} < 1$ (see, e.g., [182]). In the remainder of the present monograph, these stability conditions are assumed to hold as we restrict the attention to steady-state behavior.

2.2.2.B Preliminaries. The partial derivative $\frac{\partial}{\partial z_i} h_i(z)|_{z=1}$ of the generating function $h_i(z)$ as introduced in Property 2.2.1 represents the mean number of type-i *children* residing in queue i at the end of a visit period generated by a type-i customer present at the start of a visit to queue i (for a formal definition of children, see below). Subsequently, we define the *exhaustiveness* Φ_i of the service discipline at queue i by

$$\Phi_i = 1 - \frac{\partial}{\partial z_i} h_i(z)|_{z=1}, \qquad i = 1, 2, \ldots, N. \tag{2.9}$$

Due to stability, we have

$$0 < \Phi_i \leq 1, \qquad i = 1, 2, \ldots, N. \tag{2.10}$$

The exhaustiveness Φ_i has the following intuitively appealing interpretation: each customer present at the start of a visit to queue i will be replaced by a number of type-i customers with mean $1 - \Phi_i$. In the present subsection, we show that the exhaustiveness plays first fiddle throughout the analysis of branching-type policies, which has not been recognized - at least to this generality - before in the literature.

From the branching property each visit period V_i at queue i starting with l_i customers consists of l_i mutually independent subvisit periods T_i generated by the type-i customers present at the start of a visit to queue i. It is convenient to have an expression for the mean subvisit period $\mathbb{E}[T_i]$ in terms of Φ_i. Since $\frac{\partial}{\partial z_i} h_i(z)|_{z=1}$ equals 1 plus the expected number of type-i arrivals $\lambda_i \mathbb{E}[T_i]$ during this subvisit period minus $\mathbb{E}[T_i]/\mathbb{E}[B_i]$, which is the expected number of type-i served during this subvisit period, we can derive, after some rewriting, the following expression,

$$\mathbb{E}[T_i] = \frac{\Phi_i \mathbb{E}[B_i]}{1 - \rho_i}, \qquad i = 1, 2, \ldots, N. \tag{2.11}$$

Furthermore, (2.11) also leads to an expression for $\mathbb{E}[X_i]$, i.e., the number of type-i customers present at the start of a visit to queue i, by observing that the mean total visit period $\mathbb{E}[V_i]$ at queue i, which equals the sum of all subvisit periods, is the product of $\mathbb{E}[X_i]$ and $\mathbb{E}[T_i]$. That is, calling upon (2.7) and (2.11) yields

$$\mathbb{E}[X_i] = \frac{\mathbb{E}[V_i]}{\mathbb{E}[T_i]} = \frac{\rho_i}{1-\rho}\frac{\mathbb{E}[S]}{\mathbb{E}[T_i]} = \frac{\lambda_i}{\Phi_i}\frac{1-\rho_i}{1-\rho}\mathbb{E}[S], \qquad i = 1, 2, \ldots, N. \tag{2.12}$$

We continue with an example.

Example 2.2.2 In the present subsection, we illustrate all concepts via the exhaustive and gated disciplines.

1. In case the *exhaustive* discipline is used at queue i, we have

   $$h_i(z_1, \ldots, z_N) = \theta_i(\sum_{j \neq i} \lambda_j(1 - z_j)), \tag{2.13}$$

   where $\theta_i(\cdot)$ denotes the LST of a busy period in an $M/G/1$ queue with arrival rate λ_i and LST of the service time distribution $\beta_i(\cdot)$. The corresponding exhaustiveness reads $\Phi_i = 1$.

2. When the *gated* discipline is implemented at queue i, the function $h_i(z_1, \ldots, z_N)$ reads

   $$h_i(z_1, \ldots, z_N) = \beta_i(\sum_{j=1}^{N} \lambda_j(1 - z_j)), \tag{2.14}$$

   with exhaustiveness $\Phi_i = 1 - \rho_i$.

□

Next, define the *offspring generating function* $f(z)$ as follows

$$f(z) := (f_1(z), \ldots, f_N(z)), \tag{2.15}$$

with for $|z_j| \leq 1$, $j = 1, 2, \ldots, N$,

$$f_i(z) := h_i(z_1, \ldots, z_i, f_{i+1}(z), \ldots, f_N(z)), \qquad i = 1, 2, \ldots, N. \tag{2.16}$$

This offspring generating function represents the generating function of the joint distribution of the numbers of customers at the end of a cycle with respect to queue 1 that are *children* of a type-i customer, where a child of a customer is recursively defined as a customer that has arrived during the service time of this customer or of one of his children. Furthermore, define for $|z_j| \leq 1$, $j = 1, 2, \ldots, N$,

$$f^{(0)}(z) := z, \tag{2.17}$$
$$f^{(k)}(z) := f(f^{(k-1)}(z)), \qquad k \geq 1, \tag{2.18}$$

where $f^{(k)}(\cdot)$ represents the k^{th} generation offspring.

2.2.2.C Queue length distribution. Since we are interested in the waiting time distribution at an arbitrary queue, we focus - without loss of generality - on queue 1 in the remainder of the present subsection.

One can prove that the PGF $X(z)$ of the joint queue length distribution at a polling instant of queue 1 satisfies the following recursion (see, e.g., [182]),

$$X(z) = X\left(f(z)\right) g(z), \tag{2.19}$$

where the *immigration generating function* $g(z)$ reads

$$g(z) = \prod_{i=1}^{N} \sigma_{i+1} \left(\sum_{j=1}^{i} \lambda_j (1 - z_j) + \sum_{j=i+1}^{N} \lambda_j (1 - f_j(z)) \right). \tag{2.20}$$

Iteration of (2.19) gives us,

$$X(z) = \prod_{k=0}^{\infty} g\left(f^{(k)}(z)\right), \tag{2.21}$$

the infinite product being convergent when the stability conditions are fulfilled.

Remark 2.2.3 In systems with zero setup times (in the sequel we add a superscript 0 for that case, to distinguish its quantities from those in systems with nonzero setup times), we have to replace (2.19) by

$$X^0(z) = X^0\left(f(z)\right) - \pi^0\left(1 - g^0(z)\right), \tag{2.22}$$

where

$$g^0(z) = \sum_{j=1}^{N} \frac{\lambda_j}{\Lambda} z_j, \qquad \text{or} \qquad g^0(z) = \sum_{j=1}^{N} \frac{\lambda_j}{\Lambda} f_j(z), \tag{2.23}$$

dependent on the behavior of the server when the system becomes empty. In the first case, the server, at the moment the system becomes empty, makes a full cycle and stops right before queue 1, whereas in the second case he stops right after queue 1. Subsequently, the probability π^0 is given by

$$\pi^0 = \left[1 + \sum_{k=0}^{\infty} \left[1 - g^0\left(f^{(k)}(0)\right)\right]\right]^{-1}. \tag{2.24}$$

After iterating (2.22) we obtain

$$X^0(z) = 1 - \pi^0 \sum_{k=0}^{\infty} \left[1 - g^0\left(f^{(k)}(z)\right)\right], \tag{2.25}$$

where the infinite sum is convergent when the stability conditions are fulfilled. □

We introduce for $i = 1, 2, \ldots, N$,

$$\begin{aligned} \tilde{h}_i(z) &:= h_i(z, 1, \ldots, 1), & (2.26)\\ \tilde{f}_i^{(k)}(z) &:= f_i^{(k)}(z, 1, \ldots, 1), & (2.27)\\ \tilde{X}(z) &:= X(z, 1, \ldots, 1). & (2.28) \end{aligned}$$

Next, the LST of the waiting time distribution can be expressed as follows (see, e.g., [42]),

$$\mathbb{E}[e^{-\omega W_1}] = \mathbb{E}[e^{-\omega W_1^{M/G/1}}] \frac{\tilde{X}(\tilde{h}_1(1 - \omega/\lambda_1)) - \tilde{X}(1 - \omega/\lambda_1)}{\tilde{X}'(1)(1 - \tilde{h}_1(1))\omega/\lambda_1}, \tag{2.29}$$

where $\mathbb{E}[e^{-\omega W_1^{M/G/1}}]$ is the LST of the waiting time distribution in the corresponding isolated $M/G/1$ queue with arrival rate λ_1 and service time distribution LST $\beta_1(\cdot)$.

Remark 2.2.4 If we define

$$\tilde{X}^0(z) := X^0(z, 1, \dots, 1), \tag{2.30}$$

the counterpart of (2.29) in systems with zero setup times can be obtained, replacing $\tilde{X}(z)$ by $\tilde{X}^0(z)$. □

Repeatedly differentiating (2.29) leads to sets of linear equations, from which the moments of the waiting times and queue lengths can be obtained numerically. Alternatively, Konheim *et al.* [132] propose the so-called descendant set approach, an iterative technique that exploits the branching structure of the model by making use of the concept of descendant sets, to obtain the moments of the waiting time. Choudhury and Whitt [65] use numerical transform inversion to compute these moments, tail probabilities, transient performance measures and more.

2.2.2.D Pseudo-conservation laws. In systems with zero setup times, the principle of work conservation leads to the so-called *conservation law*, which states a certain linear relationship between the mean waiting times $\mathbb{E}[W_i^0]$ of customers of all queues which is independent of the (work-conserving) scheduling discipline (see, e.g., [47]),

$$\sum_{i=1}^{N} \rho_i \mathbb{E}[W_i^0] = \rho \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)}. \tag{2.31}$$

If setup times are nonzero, the principle of work decomposition gives rise to a so-called *pseudo-conservation law*, which is again a linear relationship among the $\mathbb{E}[W_i]$ - the affix *pseudo* is used since the resulting expression now *does* depend on the service discipline - (see again [47]),

$$\sum_{i=1}^{N} \rho_i \mathbb{E}[W_i] = \rho \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)} + \frac{\rho}{2\mathbb{E}[S]} \mathbb{V}\text{ar}[S] + \frac{\mathbb{E}[S]}{2(1-\rho)} \sum_{i=1}^{N} \rho_i(1-\rho_i) + \sum_{i=1}^{N} \mathbb{E}[M_i], \tag{2.32}$$

with M_i denoting the amount of work left behind by the server at queue i at the completion of a visit to this queue, which can be derived in closed form. That is, it is readily verified that with the help of (2.12) for $i = 1, 2, \dots, N$,

$$\mathbb{E}[M_i] = \mathbb{E}[X_i]\mathbb{E}[B_i](1-\Phi_i) = \frac{1-\Phi_i}{\Phi_i} \frac{\rho_i(1-\rho_i)}{1-\rho} \mathbb{E}[S]. \tag{2.33}$$

Thus, the term $\mathbb{E}[M_i]$ is completely determined by the service discipline at queue i and is independent of (the service discipline at) the other queues. Although this pseudo-conservation law does not give explicit expressions for the mean waiting times themselves, it appears to be a tool for developing approximations and can provide a useful check for the accuracy of simulations, numerical calculations and approximations (see, e.g., [47]). Furthermore, it gives a relatively simple expression for the weighted sum of the mean waiting times, which may be used as a first indication of overall system performance. We continue with an example.

Example 2.2.5 We now take a second look at the policies in Example 2.2.2.

1. For the exhaustive discipline at queue 1, we have $\mathbb{E}[M_1] = 0$.

2. In case the gated discipline is implemented at queue 1, one obtains $\mathbb{E}[M_1] = \frac{\rho_1^2}{1-\rho}\mathbb{E}[S]$.

□

2.2.2.E Asymptotics. As seen, the interdependence of the queueing processes in polling systems prohibits an exact explicit analysis with closed-form expressions, leading to the need of using numerical techniques to determine performance measures of interest (see, e.g., [65; 132]). However, these numerical approaches for the analysis of polling systems have several drawbacks. Firstly, numerical techniques do not reveal explicitly how the system performance depends on the system parameters and can, therefore, contribute to the understanding of the system behavior only to a limited extent. Exact closed-form expressions provide much more insight into the dependence of the performance measures on the system parameters, which leads to significant insights in the behavior of the system, e.g., insensitivity and monotonicity properties. Secondly, the efficiency of the numerical algorithms tends to degrade significantly for heavily loaded, highly asymmetric systems with a large number of queues, while the proper operation of the system is particularly critical when the system is heavily loaded. In these circumstances, one naturally resorts to asymptotic estimates. In particular, below we study the behavior of the (scaled) waiting time in case of increasing load and/or increasing setup times.

Increasing load. Van der Mei [165] studies the waiting time distribution for branching-type polling systems (and general setup time distributions with finite first two moments) under heavy traffic. That is, the waiting time distribution is considered as a function of ρ where the arrival rates are variable, while the service time distributions and the ratios of the arrival rates are fixed. This permits to parameterize the variables as a function of ρ. Subsequently, a closed-form expression for the scaled asymptotic waiting time, i.e., the limit of $1-\rho$ times the waiting time, is obtained when ρ tends to 1.

For that purpose, [165] relies on the following result on multi-type branching processes with immigration in each state [180]: the joint probability distribution of the N-dimensional branching process $\{Z_n, n = 0, 1, \ldots\}$ converges in distribution to $v\Gamma(\alpha,\mu)$ in the sense that (we take $\xrightarrow{d}$ to represent convergence in distribution),

$$\lim_{n\to\infty} \frac{1}{\pi_n(\xi)} Z_n \xrightarrow{d} v\Gamma(\alpha,\mu), \qquad \xi \uparrow 1, \tag{2.34}$$

where ξ is the maximum eigenvalue of the so-called mean matrix, $\pi_n(\xi)$ is a scaling function, v is a known N-dimensional vector and $\Gamma(\alpha,\mu)$ is a gamma-distributed random variable with known shape and scale parameters α and μ, respectively. Thereupon, in [165] it is shown that this result leads to asymptotic heavy-traffic results for the (joint) queue length distributions at polling instants in branching-type polling systems. We have seen before that such results readily lead to corresponding results for the waiting time distributions as summarized in the theorem below.

Theorem 2.2.6 *The LST of the distribution of the asymptotic scaled waiting time under heavy traffic is given by*

$$\begin{aligned}\mathbb{E}[e^{-\omega(1-\rho)W_1}] \;\to\; & \frac{1}{(1-\rho_1)\mathbb{E}[S]\omega}\Big[\left(\frac{\delta\beta\Phi_1}{\delta\beta\Phi_1+(1-\Phi_1)(1-\rho_1)\omega}\right)^{\beta\delta\mathbb{E}[S]} - \\ & \left(\frac{\delta\beta\Phi_1}{\delta\beta\Phi_1+(1-\rho_1)\omega}\right)^{\beta\delta\mathbb{E}[S]}\Big], \qquad \rho\uparrow 1, \end{aligned}\tag{2.35}$$

where

$$\beta = \frac{\mathbb{E}[B]}{\mathbb{E}[B^2]}, \qquad \textit{and} \qquad \delta = \sum_{i=1}^{N}\left(\frac{\rho_i(1-\rho_i)(1-\Phi_i)}{\Phi_i} + \rho_i\sum_{j=i+1}^{N}\rho_j\right). \tag{2.36}$$

Proof. See [165]. □

We continue with an example.

Example 2.2.7 For the policies introduced in Example 2.2.2, Theorem 2.2.6 has the following implications.

1. In case of the exhaustive discipline at queue 1, one obtains

$$\mathbb{E}[e^{-\omega(1-\rho)W_1}] \to \frac{1}{(1-\rho_1)\mathbb{E}[S]\omega}\left[1-\left(\frac{\delta\beta}{\delta\beta+(1-\rho_1)\omega}\right)^{\beta\delta\mathbb{E}[S]}\right], \qquad \rho\uparrow 1. \quad (2.37)$$

2. Implementing the gated discipline at queue 1 yields

$$\mathbb{E}[e^{-\omega(1-\rho)W_1}] \to \frac{1}{(1-\rho_1)\mathbb{E}[S]\omega}\left[\left(\frac{\delta\beta}{\delta\beta+\rho_1\omega}\right)^{\beta\delta\mathbb{E}[S]}-\left(\frac{\delta\beta}{\delta\beta+\omega}\right)^{\beta\delta\mathbb{E}[S]}\right], \qquad \rho\uparrow 1. \quad (2.38)$$

□

Increasing setup times. Winands [P11] presents an exact asymptotic analysis of the waiting time distribution in branching-type polling systems with *deterministic* setup times when the setup times tend to infinity. Since the waiting time grows to infinity in the limiting case, [P11] focuses on the asymptotic scaled waiting time $W_1/\mathbb{E}[S]$ as $\mathbb{E}[S]\to\infty$, where the ratios of the setup times remain constant. In order to derive these asymptotics, [P11] builds upon a result of [42] which derives a strong relation between the waiting time distributions in models *with* and *without* setup times. This relation is established by relating the similarities in the offspring generating functions of the underlying branching processes and by expressing the differences between the underlying immigration functions.

In particular, [42] shows that the LST of the waiting time distribution of a type-1 customer is given by

$$\mathbb{E}[e^{-\omega W_1}] = \mathbb{E}[e^{-\omega W_1^0}]\frac{e^{-\mathbb{E}[S]\tilde{H}(\tilde{h}_1(1-\omega/\lambda_1))}-e^{-\mathbb{E}[S]\tilde{H}(1-\omega/\lambda_1)}}{\mathbb{E}[S][\tilde{H}(1-\omega/\lambda_1)-\tilde{H}(\tilde{h}_1(1-\omega/\lambda_1))]}, \quad (2.39)$$

where W_1^0 is the waiting time in the corresponding polling system with zero setup times. At this point, we feel it is worth reminding the reader that no closed-form expression for $\mathbb{E}[e^{-\omega W_1^0}]$ is known. Furthermore, $\tilde{H}(y)$ is defined as

$$\tilde{H}(y) := \sum_{k=0}^{\infty}\sum_{i=1}^{N}\lambda_i(1-\tilde{f}_i^{(k)}(y)). \quad (2.40)$$

Since the mean number of type-1 customers present at the start of a visit to queue 1 is exactly equal to the average offspring of customers which arrived during a setup time, we have the following relation between $\tilde{H}'(1)$ and $\mathbb{E}[X_1]$,

$$\mathbb{E}[X_1] = -\mathbb{E}[S]\tilde{H}'(1), \quad (2.41)$$

and, thus, by applying (2.12),

$$\tilde{H}'(1) = -\frac{\lambda_1}{\Phi_1}\frac{1-\rho_1}{1-\rho}. \quad (2.42)$$

The decomposition as expressed in (2.39) can be used to derive an explicit expression for the LST of the distribution of the asymptotic scaled waiting time as presented in the lemma below.

Lemma 2.2.8 *In case of deterministic setup times, the LST of the distribution of the asymptotic scaled waiting time is given by*

$$\mathbb{E}[e^{-\omega \frac{W_1}{\mathbb{E}[S]}}] \rightarrow \frac{1-\rho}{(1-\rho_1)\omega}\left(e^{-\frac{1-\Phi_1}{\Phi_1}\frac{1-\rho_1}{1-\rho}\omega} - e^{-\frac{1}{\Phi_1}\frac{1-\rho_1}{1-\rho}\omega}\right), \qquad (\mathbb{E}[S] \rightarrow \infty). \tag{2.43}$$

Proof. First of all, the term $\mathbb{E}[e^{-\omega W_1^0}]$ in (2.39) does not depend on S implying that

$$\mathbb{E}[e^{-\omega \frac{W_1^0}{\mathbb{E}[S]}}] \rightarrow 1, \qquad (\mathbb{E}[S] \rightarrow \infty). \tag{2.44}$$

Next, we observe that

$$\mathbb{E}[S]\tilde{H}(1-\frac{\omega}{\lambda_1\mathbb{E}[S]}) = -\frac{\omega}{\lambda_1}\frac{\tilde{H}(1-\frac{\omega}{\lambda_1\mathbb{E}[S]})-\tilde{H}(1)}{-\frac{\omega}{\lambda_1\mathbb{E}[S]}} \xrightarrow{\mathbb{E}[S]\rightarrow\infty} -\frac{\omega}{\lambda_1}\tilde{H}'(1) = \frac{1}{\Phi_1}\frac{1-\rho_1}{1-\rho}\omega, \tag{2.45}$$

where the first equation follows from the fact that $\tilde{H}(1) = 0$ and the last equation from (2.42). Similarly, we have that

$$\begin{aligned}\mathbb{E}[S]\tilde{H}(\tilde{h}_1(1-\frac{\omega}{\lambda_1\mathbb{E}[S]})) &= -\frac{\omega}{\lambda_1}\frac{\tilde{H}(\tilde{h}_1(1-\frac{\omega}{\lambda_1\mathbb{E}[S]}))-\tilde{H}(\tilde{h}_1(1))}{-\frac{\omega}{\lambda_1\mathbb{E}[S]}} \\ &\xrightarrow{\mathbb{E}[S]\rightarrow\infty} -\frac{\omega}{\lambda_1}\tilde{H}'(\tilde{h}_1(1))\tilde{h}_1'(1) \\ &= \frac{1-\Phi_1}{\Phi_1}\frac{1-\rho_1}{1-\rho}\omega, \end{aligned} \tag{2.46}$$

where the last equality follows from the definition of the exhaustiveness factor and (2.42). Substituting (2.45) and (2.46) into (2.39) completes the proof (after some rewriting). □

Since the righthand side of (2.43) is recognized as the LST of the uniform distribution, Lemma 2.2.8 leads to the following result for the distribution of the asymptotic scaled waiting time.

Theorem 2.2.9 *In case of deterministic setup times, the distribution of the asymptotic scaled waiting time is given by*

$$\frac{W_1}{\mathbb{E}[S]} \xrightarrow{d} \frac{1-\rho_1}{1-\rho}U_1, \qquad (\mathbb{E}[S] \rightarrow \infty), \tag{2.47}$$

where U_1 is uniformly distributed on $[\frac{1-\Phi_1}{\Phi_1}, \frac{1}{\Phi_1}]$.

Proof. Follows directly from Lemma 2.2.8 in combination with the convergence theorem of Feller for LSTs (see, e.g., p. 652 of [68]). □

Below we continue our example.

Example 2.2.10 Let us return to the policies introduced in Example 2.2.2.

1. In case of the exhaustive discipline at queue 1, the scaled waiting time, as the setup times tend to infinity, is uniformly distributed on $[0, \frac{1-\rho_1}{1-\rho}]$.
2. When the gated discipline is implemented at queue 1, the scaled waiting time for increasing setup times follows a uniform distribution on $[\frac{\rho_1}{1-\rho}, \frac{1}{1-\rho}]$.

□

Increasing load and increasing setup times. If we take a closer look at (2.35) we see that in heavy traffic the impact of higher moments of the setup times on the waiting time distribution vanishes, i.e., the scaled asymptotic waiting time depends on the marginal setup time distributions only through the first moment of the total setup time in a cycle. Building upon this observation, Winands [P11] studies the scaled asymptotic waiting time in branching-type polling systems with *generally distributed setups under heavy traffic* when the setup times tend to infinity. The only restriction made on the setup times is that the first two moments of all the setup times should exist, i.e., they should be finite. Firstly, [P11] lets the arrival rates increase in such a way that ρ tends to 1 (while the service time distributions and the ratios of the arrival rates are fixed), which allows to exploit the heavy-traffic results from [165]. Secondly, [P11] lets the mean total setup time in a cycle $\mathbb{E}[S]$ tend to infinity. This step-by-step plan is formalized in the following lemma.

Lemma 2.2.11 *In case of general setup times, the LST of the distribution of the asymptotic scaled waiting time under heavy traffic is given by*

$$\mathbb{E}[e^{-\omega(1-\rho)\frac{W_1}{\mathbb{E}[S]}}] \;\to\; \frac{1}{(1-\rho_1)\omega}\left(e^{-\frac{1-\Phi_1}{\Phi_1}(1-\rho_1)\omega} - e^{-\frac{1}{\Phi_1}(1-\rho_1)\omega}\right), \qquad (\rho \uparrow 1 \quad \textit{and then} \quad \mathbb{E}[S] \to \infty). \tag{2.48}$$

Proof. First of all, we let ρ tend to 1 in such a way that we can apply the limit theorems of [165], which imply that (2.35) holds. Applying the following standard limit result,

$$\lim_{x\to\infty}\left(\frac{a}{a+\frac{b}{x}}\right)^{cx} = e^{-\frac{bc}{a}}, \tag{2.49}$$

to the scaled waiting time $(1-\rho)\frac{W_1}{\mathbb{E}[S]}$ in (2.35) completes the proof (after some straightforward manipulations). □

Lemma 2.2.11 has the following immediate consequence.

Theorem 2.2.12 *In case of general setup times, the distribution of the asymptotic scaled waiting time under heavy traffic is given by*

$$\frac{(1-\rho)W_1}{\mathbb{E}[S]} \xrightarrow{d} (1-\rho_1)U_1, \qquad (\rho \uparrow 1 \quad \textit{and then} \quad \mathbb{E}[S] \to \infty), \tag{2.50}$$

where U_1 is uniformly distributed on $[\frac{1-\Phi_1}{\Phi_1}, \frac{1}{\Phi_1}]$.

Proof. Follows directly from Lemma 2.2.11 in combination with the convergence theorem of Feller for LSTs (see, e.g., p. 652 of [68]). □

We note that in the case of deterministic setup times Theorem 2.2.12 with "$\rho \uparrow 1$ and then $\mathbb{E}[S] \to \infty$" replaced by "$\mathbb{E}[S] \to \infty$ and then $\rho \uparrow 1$" is implied by Theorem 2.2.9 and, subsequently, letting ρ tend to 1. This implies that also in this limiting regime the scaled waiting time is uniformly distributed on $[\frac{1-\Phi_1}{\Phi_1}, \frac{1}{\Phi_1}]$.

Winands [P11] extends Theorems 2.2.9 and 2.2.12 in various directions: to queue length distributions at polling instants, to joint distributions of the queue lengths and to systems with the globally-gated service policy [52] (which does not satisfy Property 2.2.1). Furthermore, [P11] suggests simple closed-form expressions for the waiting time distributions in systems with finite setup times and provides conjectures for the asymptotic behavior of systems with renewal arrivals. The interested reader is referred to [P11] for more details. Finally, we return to our example.

Example 2.2.13 For the final time, we return to the policies introduced in Example 2.2.2.

1. In case of the exhaustive discipline at queue 1, the scaled waiting time under increasing setup times in heavy traffic is uniformly distributed on $[0, 1-\rho_1]$.
2. When the gated discipline is implemented at queue 1, the scaled waiting time, when setup times tend to infinity, in heavy traffic follows a uniform distribution on $[\rho_1, 1]$.

□

We close this subsection with a remark.

Remark 2.2.14 For general traffic settings we have seen that all moments, and in particular the standard deviation, of the queue length in branching-type polling systems can be obtained numerically (see, e.g., [65; 132]). However, these procedures are complicated, time-consuming and lack transparency, which raises the importance of *explicit*, either exact or approximate, expressions for the standard deviation. Explicit *exact* formulae for the standard deviation have only been derived in systems with two queues *without* setup times [201] and *with* setup times [83], in symmetric systems with two, three or four queues [137] and symmetric continuous polling systems with deterministic setup times [95], i.e., systems consisting of an infinite number of queues. Explicit *approximate* expressions for the standard deviations are scarce as well; only approximations for heavily loaded systems and models with large setup times have been derived so far. Although such approximations can be used in some specific traffic situations, their accuracy worsens for decreasing load and decreasing setup times, respectively. Motivated by these challenges, Winands [P12] develops a novel closed-form approximation, which is accurate over the entire range of parameters, by using results for symmetric continuous polling systems in conjunction with heavy-traffic results. Support for the quality of the approximation is, first of all, provided by results of an extensive numerical evaluation. Second, in some specific systems [P12] analytically derives bounds for the standard deviations based on the approximation, which are proven to be tight. Finally, [P12] proves that the approximation is in line with existing results for the aforementioned extreme cases, i.e., for heavily loaded systems, for systems with large setup times and for systems with an infinite number of queues. □

2.2.3 k-limited policy

In the present subsection, we study the k-limited service discipline which does not satisfy Property 2.2.1. In contrast to branching-type policies, for k-limited policies no exact general framework exists which reveals itself in the limited results available in the literature. Below we discuss these scarce results.

2.2.3.A Stability. For the k-limited discipline, a necessary and sufficient stability condition reads (see [103] for a rigorous proof),

$$\rho + \mathbb{E}[S] \max_{1,2,\ldots,N} \frac{\lambda_i}{k_i} < 1. \tag{2.51}$$

If the system is stable, (2.51) may be rewritten by using (2.5) as follows, for $i = 1, 2, \ldots, N$,

$$\lambda_i \mathbb{E}[C] < k_i. \tag{2.52}$$

In words, this means that for a stable system the average number of type-i customers arriving in a cycle is smaller than the service limit k_i, i.e., the maximum number of type-i customers served in a cycle. Throughout the assumption is made that (2.51) is fulfilled.

2.2.3.B Queue length distribution. To this very day, not only hardly any *exact* results for the queue length distribution, or even the means, in polling systems with the k-limited service policy have been obtained, but also their derivations give little hope for extensions to more realistic systems. That is, Groenendijk [116] and Ibe [122] give an explicit LST for the waiting time distribution in a two-queue 1-limited/exhaustive system. For two-queue systems where both queues are served according to the 1-limited discipline, the problem of finding the queue length distribution can be shown to translate into a boundary value problem [46; 49; 70; 84]. Fuhrmann [105] obtains the mean queue length in symmetric 1-limited polling systems with an arbitrary number of queues. For general k, an exact evaluation for the queue length distribution is only available for very few special two-queue cases (see Lee [145] and Ozawa [176; 177]). In these models one (low-priority) queue is served by the k-limited service strategy, whereas the other (high-priority) queue is served by the exhaustive policy. Furthermore, [145; 176; 177] make the restrictive assumption of *zero setup times*. Although the complexity of exhaustive and gated systems typically go hand in hand, the exception that proves the rule is the two-queue gated/k-limited model that, even in the case the service limit equals 1, appears to defy an exact analysis. In the absence of exact results for the marginal queue length distributions, people have resorted to *numerical* approaches, such as the power series algorithm [41] and techniques based on discrete Fourier transforms [147]. The main disadvantage of both methods is that time and memory requirements are exponential functions of the number of queues.

2.2.3.C Pseudo-conservation laws. In case of the k-limited policy, the general form of the pseudo-conservation law as shown in (2.31) still holds. However, the unknowns $\mathbb{E}[M_i]$ form the stumbling block to the straightforward application of this law. That is, if we denote by Y_i the number of customers served at queue i in a cycle, we have

$$\mathbb{E}[M_i] = \left(\mathbb{E}[X_i] - (1-\rho_i)\mathbb{E}[Y_i]\right)\mathbb{E}[B_i], \tag{2.53}$$

where, due to stability, $\mathbb{E}[Y_i] = \lambda_i \frac{\mathbb{E}[S]}{1-\rho}$. In contrast to branching-type policies, for the k-limited policy no exact results for $\mathbb{E}[X_i]$ are known. Everitt [88] expresses these unknown quantities in terms of $\mathbb{E}[Y_i^{(2)}] = \mathbb{E}[Y_i(Y_i-1)]$, the second factorial moments of the number of customers served at queue i per cycle, for $i = 1, 2, \ldots, N$,

$$\mathbb{E}[X_i] = \frac{1}{k_i}\left(\lambda_i \mathbb{E}[Y_i]\mathbb{E}[W_i] - \frac{1-\rho_i}{2}\mathbb{E}[Y_i^{(2)}] + \rho_i \mathbb{E}[Y_i]\right) + (1-\rho_i)\mathbb{E}[Y_i], \tag{2.54}$$

which yields,

$$\begin{aligned}\sum_{i=1}^{N} \rho_i \left(1 - \frac{\mathbb{E}[Y_i]}{k_i}\right)\mathbb{E}[W_i] \;=\;& \rho \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)} + \frac{\rho}{2\mathbb{E}[S]}\mathbb{V}\mathrm{ar}[S] + \frac{\mathbb{E}[S]}{2(1-\rho)}\sum_{i=1}^{N}\rho_i(1-\rho_i) + \\ & \frac{\mathbb{E}[S]}{1-\rho}\sum_{i=1}^{N}\frac{\rho_i^2}{k_i} - \sum_{i=1}^{N}\frac{\rho_i(1-\rho_i)}{2}\frac{\mathbb{E}[Y_i^{(2)}]}{\lambda_i k_i}.\end{aligned} \tag{2.55}$$

For the special case that all service limits k_i equal 1, we can compute $\mathbb{E}[Y_i^{(2)}]$ explicitly. That is, we know that one service is completed per cycle with probability $\mathbb{E}[Y_i]$ and no service with probability $1 - \mathbb{E}[Y_i]$ which implies that $\mathbb{E}[Y_i^{(2)}] = 0$ and, thus,

$$\mathbb{E}[M_i] = \rho_i \mathbb{E}[Y_i] \left(\mathbb{E}[W_i] + \mathbb{E}[B_i]\right). \tag{2.56}$$

In case of general k_i, we have to resort to approximations of and bounds on the unknown terms $\mathbb{E}[Y_i^{(2)}]$, for which we refer to [59; 60; 88] and the references therein.

2.2.3.D Asymptotics. Besides the paper of Lee [145] for a two-queue model, no studies have appeared so far analyzing the asymptotic behavior, due to either increasing load or increasing setup times, of the k-limited discipline in an exact setting. More specifically, [145] investigates a two-queue model without setup times, where queue 1 is served exhaustively and queue 2 is served according to the k-limited policy. Now, we fix λ_2 and increase λ_1 in such a way that ρ tends to one. In this limit queue 1 remains stable, i.e., it behaves like an $M/G/1$ system with vacations with arrival rate λ_1, service time LST $\beta_1(\cdot)$ and vacation time LST $\beta_2^k(\cdot)$. Queue 2, however, becomes unstable in the limit and the scaled amount of work in this queue equals the scaled amount of work in an $M/G/1$ queue in which two customer classes are combined into one customer class with arrival rate $\lambda_1 + \lambda_2$ and service times with LST $\frac{\lambda_1}{\lambda_1+\lambda_2}\beta_1(\cdot) + \frac{\lambda_2}{\lambda_1+\lambda_2}\beta_2(\cdot)$. The main result of [145] is that in the limit the number of customers in queue 1 and the scaled amount of work in queue 2 become independent.

2.2.4 Contributions of the present monograph

Chapter 1 has already positioned our contributions in the area of production-inventory systems. Now, we relate these contributions to the field of polling systems both to the class of branching-type policies and the k-limited policy. We stress the one-to-one correspondence between this distinction and the research objectives of Chapter 1.

Branching-type policies. As elaborated on in the preceding subsection, the analysis of branching-type policies is a well-trodden area in terms of results, but it is relatively unexplored in terms of methodology. That is, results are not obtained via a single unifying approach and part of the results is obtained by applying deep theorems from multi-type branching processes. These observations have motivated us to develop a unifying framework, the so-called *Mean Value Analysis* (MVA) framework, for the most important representatives of the branching class (the exhaustive and gated discipline). By confining ourselves to *average* performance measures, we show that all results discussed, i.e., marginal waiting times, pseudo-conservation laws and asymptotics, can by obtained by calling only upon two basic queueing results: the PASTA property, i.e., Poisson arrivals see time averages [220] and Little's Law [154]. Finally, the MVA framework allows us to analyze a myriad of scheduling disciplines, besides the traditional FCFS policy, implying that we can solve the scheduling decision to optimality.

k-limited policy. For the k-limited discipline, we have observed a paucity of results. We present the first exact analysis of a two-queue k-limited/exhaustive system *with* (state-dependent) setup times and we indicate the difficulties in extending the analysis to a more general system. Furthermore, we develop an accurate, robust and computationally efficient approximate algorithm for the evaluation of k-limited polling systems with a general number of queues and generally distributed interarrival, setup and service times. The methodology used in this algorithm goes beyond the k-limited policy and, therefore, we can look upon this algorithm as a pilot study for the approximation of a general class of polling systems. Finally, we present a simulation study, which examines the quality of the k-limited policy as priority mechanism and gives some explorative results for the optimization of the service limits.

Chapter 3

MVA framework

In the present chapter we develop a *mean value analysis* (MVA) framework for the computation of mean waiting times in exhaustive and gated polling systems. We start the development of this MVA framework in Section 3.1 with the computation of the mean marginal waiting times and the derivation of the corresponding pseudo-conservation law for general traffic settings. Next, we extend the framework to the asymptotic regime of high utilization of capacity due to either high load or large setup times in Sections 3.2 and 3.3, respectively. Subsequently, we show that within the MVA framework examination of various scheduling disciplines is possible (see Section 3.4). Finally, the chapter is wound up in Section 3.5 with conclusions and a list of possible extensions.

3.1 General traffic settings

The present section, which is an abridged version of the papers [P15; P16; P17], is concerned with the exhaustive and gated service disciplines. The single most important performance measure for polling systems is, in many applications, the mean waiting time of a customer. Unfortunately, explicit closed-form expressions for the mean waiting times in polling systems with exhaustive-type or gated-type service are only known in very special cases (see, also, Chapter 2). We develop a novel approach to compute the mean waiting times in general continuous-time polling systems with either exhaustive or gated service, the so-called MVA.

In the past several other approaches have been proposed for computing these mean waiting times, of which some extend to the computation of higher moments as well. One such method is the *buffer occupancy* method as developed by [71; 72; 83], which is closely related to the classical generating function approach as expounded in Chapter 2. This method is based on the buffer occupancy variables $X_{i,j}$, which denote the queue length at queue j at a polling instant of queue i, $i, j = 1, 2, \ldots, N$. The buffer occupancy method requires the solution of N^3 linear equations with unknowns $\mathbb{E}[X_{i,j}X_{i,k}]$ to compute the mean waiting times in *all* N stations simultaneously. These equations may be efficiently solved in an iterative manner requiring $O(N^3 \log_\rho \epsilon)$ operations (additions and multiplications), where ϵ is the relative accuracy required (see [150]). Based on this buffer occupancy method, [132] developed the *descendant set* method; an iterative technique that computes the mean waiting time at each queue independently of the other queues. The descendant set approach is based on counting the number of descendants of each customer in the system. This method requires $O(N \log_\rho \epsilon)$ operations for the computation of the mean waiting time in a *single* station. A second well-known method based on the buffer occupancy method is the *individual station* technique [197], which also allows, as the name suggests, the individual computation of the mean waiting time at each queue. The individual station technique is, however, not an iterative approach. The mean waiting time at a *single* queue is computed in $O(N^2)$ operations, which obviously does not depend on the system utilization contrary to the computational complexities of the aforementioned methods.

Besides the techniques based on the buffer occupancy method, another school of approaches is the one embroidering on the *station time* method [96]. In the station time approach, all mean waiting times are obtained simultaneously starting from the station time variables θ_i, $i = 1, 2, \ldots, N$. The station time θ_i is composed of the time the server spends servicing customers at queue i plus the *preceding* setup time in case of exhaustive service or plus the *succeeding* setup time in case of gated service. The station time technique induces a set of N^2 linear equations with unknowns $\mathbb{E}[\theta_i \theta_j]$, which can be solved iteratively in $O(N^2 \log_\rho \epsilon)$ operations leading to *all* N mean waiting times. An extension of the station time method is the approach developed by [187]. Their approach induces a set of only N linear equations, which is, however, less sparse. Solving this set of equations requires $O(N^3)$ operations for *all* N waiting time figures.

Recently, *Hirayama et al.* [120] developed a third alternative method for obtaining the mean waiting times. The authors analyze first the mean waiting times conditioned on the state of the system at an arrival epoch. Then, from the analysis of the system at polling instants, a set of linear functional equations for these conditional waiting times is obtained. By applying a limiting procedure, they derive a set of $N(N+1)$ linear equations for the unconditional mean waiting times, which can be solved in $O(N^6)$ operations. With respect to this computational complexity, it should be noted that the method of [120] shows some similarities with the buffer occupancy approach and that it, therefore, may be possible to construct more efficient iterative algorithms to solve their set of equations.

With respect to the above-mentioned methods, two issues are noteworthy. Firstly, each of the approaches can be readily adapted to a discrete-time counterpart, apart from some occasional subtleties (see, e.g., [133; 184; 200]). The second important observation is that when comparing the use of the aforementioned approaches in the open literature over the recent years, it immediately strikes the eye that the buffer occupancy method and its variations can be - or at least have been - applied to the widest variety of polling systems. In fact, this method appears to be applicable to the complete class of service disciplines satisfying the branching property (see Chapter 2). However, the techniques based on the station time method and Hirayama's method have been applied to a restricted class of polling systems only. For example, it is known that the station time method cannot be used in polling systems with *mixed service*, where some of the queues are served according to the exhaustive policy and some by the gated strategy.

The objective of the present section is the development of a novel approach to compute the mean waiting times for exhaustive-type or gated-type polling systems in a purely probabilistic manner. More specifically, we derive a set of N^2 and $N(N+1)$ linear equations for these waiting time figures in case of exhaustive and gated service, respectively, with the help of the following two basic queueing results: (i) the PASTA property, i.e., Poisson arrivals see time averages [220] and (ii) Little's Law [154]. The unknowns in these equations are $\mathbb{E}[Q_{i,j}]$, the mean queue length at queue i at an arbitrary epoch within a station time of queue j. The method of the present section can be looked upon as a MVA for general polling systems with exhaustive or gated service. MVA is known as a powerful tool to determine mean performance measures in all kinds of queueing models (see, e.g., [142]), but it has never been applied to polling systems.

The main contribution of the present section is two-fold. The first main contribution can be found in the set of equations itself. In contrast to most of the above-mentioned approaches, the unknowns in these equations are all *first* moments of (residual) random variables and, thus, no correlation terms are required. Furthermore, MVA evaluates the polling system at arbitrary epochs in time and not on embedded points such as polling instants. An additional merit of MVA is the fact that it allows for an evaluation of polling systems with mixed service. Finally, MVA results in the solution of no more than N^2 and $N(N+1)$ linear equations for *all* N mean waiting times in case of exhaustive and gated service, respectively. In the past, various efficient algorithms have been developed based on the buffer occupancy method and the station time approach, which require systems of up to N^3 equations. It may, therefore, be possible to construct just as efficient iterative algorithms based on MVA. We emphasize that the development of such an algorithm is not within the scope of the present monograph.

The second main contribution, which is perhaps even more important, lies in the derivation of the set of equations by MVA. This derivation is based on standard queueing results and has a probabilistic interpretation all the way. Consequently, it is rather straightforward to apply MVA to variants of the considered polling systems: (i) systems with Poisson batch arrivals, (ii) systems with fixed polling tables and (iii) discrete-time polling systems. Finally, MVA may open new ways for the evaluation, both in an exact and approximate manner, of other polling systems.

In the course of our analysis, we obtain short proofs of three seminal results for (exhaustive) polling systems as by-product. First, we present an elementary proof of the *variance absorption* result (see, e.g., [187; 73]), which states that an interchange exists among the variance in service and setup times such that the mean waiting times remain unchanged. Second, it turns out that a by-product of MVA is a simple proof of the *decomposition result*, which establishes that the mean waiting times in polling systems without setup times and systems with deterministic setup times differ only by simple constants (see, e.g., [106; 198]). Third, we obtain a short derivation of the *pseudo-conservation law* for exhaustive polling systems (see [47]), which gives a linear relation between the mean waiting times of customers of all queues.

The structure of the present section is as follows. First, Subsection 3.1.1 introduces further notation. Subsection 3.1.2 presents the main result of the section: the derivation of a set of equations for the mean waiting time in polling systems with exhaustive service. In Subsection 3.1.3 we build upon these results to obtain the pseudo-conservation law for the system under consideration. As a by-product MVA computes second-order moments of the station times and, in particular, the correlation between successive station times as elucidated in Subsection 3.1.4. In Subsection 3.1.5, it is shown that the application of MVA to systems with gated service, with mixed service or with fully gated service is rather straightforward.

3.1.1 Notation

The system of interest is the basic N-queue polling system as introduced in Chapter 2 with service at each queue either according to the exhaustive or gated discipline (for the case of mixed service we refer to Subsection 3.1.5). For presentation reasons, we first focus on the case $\mathbb{E}[S] > 0$. When the total setup time is equal to zero, some subtleties appear due to the fact that the number of cycles with zero length tends to infinity. The station time θ_i of queue i, $i = 1, 2, \dots, N$, is composed of the service period of queue i, the time the server spends servicing customers at queue i, plus the *preceding* setup time in case of exhaustive service or plus the *succeeding* setup time in case of gated service. By virtue of these two different definitions, a queue is empty exactly at the end of its station time in case of exhaustive service, while the queue before the gate is empty at the beginning of a station time in case of gated service (all customers waiting for service are then placed behind the gate).

Since the server is working a fraction ρ_i of the time on queue i, the mean of a station period of queue i reads, for exhaustive service,

$$\mathbb{E}[\theta_i] = \rho_i \mathbb{E}[C] + \mathbb{E}[S_i], \qquad i = 1, 2, \dots, N, \tag{3.1}$$

and, for gated service,

$$\mathbb{E}[\theta_i] = \rho_i\mathbb{E}[C] + \mathbb{E}[S_{i+1}], \qquad i = 1, 2, \ldots, N. \tag{3.2}$$

We define an (i, j)-period $\theta_{i,j}$ as the sum of j consecutive station times starting in queue i, for $i, j = 1, 2, \ldots, N$. The corresponding mean is given by

$$\mathbb{E}[\theta_{i,j}] = \sum_{n=i}^{i+j-1} \mathbb{E}[\theta_n], \qquad i = 1, 2, \ldots, N, \qquad j = 1, 2, \ldots, N. \tag{3.3}$$

Notice that in case $j = 1$ and $j = N$, $\mathbb{E}[\theta_{i,j}]$ is equal to the mean station period $\mathbb{E}[\theta_i]$ of queue i and the mean cycle length $\mathbb{E}[C]$, respectively.

The fraction of time $q_{i,j}$ the system is in an (i, j)-period equals

$$q_{i,j} = \frac{\mathbb{E}[\theta_{i,j}]}{\mathbb{E}[C]}, \qquad i = 1, 2, \ldots, N, \qquad j = 1, 2, \ldots, N, \tag{3.4}$$

where, by definition, $q_{i,N}$ equals 1. Moreover, the mean of a residual (i, j)-period is given by

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{\mathbb{E}[\theta_{i,j}^2]}{2\mathbb{E}[\theta_{i,j}]}, \qquad i = 1, 2, \ldots, N, \qquad j = 1, 2, \ldots, N, \tag{3.5}$$

with the remark that the second moments $\mathbb{E}[\theta_{i,j}^2]$ are still unknown at this stage. Notice that since for each fixed (i, j) the successive (i, j)-periods are not independent, they do not form a renewal process. This means, among others, that (3.5) does not directly follow from the theory of regenerative processes. For a proof why this result is nevertheless still valid see, e.g., [99].

Our main interest is in the mean waiting time $\mathbb{E}[W_i]$ of a type-i customer, $i = 1, 2, \ldots, N$. By Little's Law, these mean waiting times are obviously related to the mean queue lengths (excluding the customer possibly in service) $\mathbb{E}[Q_i]$, $i = 1, 2, \ldots, N$. The analysis of the present section is oriented towards the determination of $\mathbb{E}[Q_{i,n}]$, the mean queue length at queue i at an arbitrary epoch within a station time of queue n, $i, n = 1, 2, \ldots, N$. The corresponding unconditional mean queue length $\mathbb{E}[Q_i]$ can be expressed in terms of $\mathbb{E}[Q_{i,n}]$ as follows

$$\mathbb{E}[Q_i] = \sum_{n=1}^{N} q_{n,1}\mathbb{E}[Q_{i,n}], \qquad i = 1, 2, \ldots, N. \tag{3.6}$$

3.1.2 Mean value analysis

The goal of the present section is the derivation of the mean waiting times in exhaustive polling systems by using MVA. This typically starts with the derivation of a so-called *arrival relation* with the help of the PASTA property. Therefore, consider a tagged customer at the moment he arrives at queue i, $i = 1, 2, \ldots, N$. Based on PASTA, we know that the state distribution seen by this tagged customer is identical to the equilibrium distribution. That is, this customer has to wait for the servicing of all customers Q_i, who were already waiting in this queue upon his arrival. Further, with probability ρ_i the server is working at queue i upon his arrival and the tagged customer has to wait for the residual service time of the customer in service as well. On the other hand, with probability $\mathbb{E}[S_i]/\mathbb{E}[C]$ the server is in a setup phase for queue i and the waiting time of the customer is increased by a residual setup time. Finally, with probability $1 - q_{i,1}$ the server is at one of the other queues and the service of the tagged customer is delayed until the server starts service again at queue i. The latter time period is obviously equal to the sum of a residual $(i + 1, N - 1)$-period and a setup time for queue i.

Hence, we have the following arrival relation for the mean waiting time $\mathbb{E}[W_i]$ of a type-i customer, $i = 1, 2, \ldots, N$,

$$\mathbb{E}[W_i] = \mathbb{E}[Q_i]\mathbb{E}[B_i] + \rho_i \mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})\Big(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]\Big). \quad (3.7)$$

Application of Little's Law, stating that $\mathbb{E}[Q_i] = \lambda_i \mathbb{E}[W_i]$, yields

$$\mathbb{E}[Q_i] = \frac{\lambda_i}{1 - \rho_i}\left(\rho_i \mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})\Big(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]\Big)\right). \quad (3.8)$$

Summarizing, we can say that (3.8) has been derived by a standard application of MVA, i.e., combining the arrival relation with Little's Law. However, the unknowns $\mathbb{E}[R_{\theta_{i+1,N-1}}]$ form the stumbling block to the straightforward computation of the mean queue lengths via this equation. To obtain the unknowns $\mathbb{E}[R_{\theta_{i+1,N-1}}]$, we relate them to $\mathbb{E}[Q_{i,n}]$ and derive a set of equations for these quantities.

Firstly, as under the exhaustive policy no type-i customers are left at the end of a station time of queue i, the following property can be obtained. The number of type-i customers present at an arbitrary moment within an $(i + 1, j)$-period equals the number of Poisson arrivals during the age of this $(i + 1, j)$-period. Since the age is in distribution equal to the residual time, the following equation holds

$$\sum_{n=i+1}^{i+j} \frac{q_{n,1}}{q_{i+1,j}}\mathbb{E}[Q_{i,n}] = \lambda_i \mathbb{E}[R_{\theta_{i+1,j}}], \qquad i = 1, 2, \ldots, N, \quad j = 1, \ldots, N - 1. \quad (3.9)$$

Secondly, substitution of (3.6) into (3.8) yields, for $i = 1, 2, \ldots, N$,

$$\sum_{n=1}^{N} q_{n,1}\mathbb{E}[Q_{i,n}] = \frac{\lambda_i}{1 - \rho_i}\left(\rho_i \mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})\Big(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]\Big)\right). \quad (3.10)$$

It is easily seen that (3.9) and (3.10) represent a set of N^2 linear equations for the unknowns $\mathbb{E}[Q_{i,n}]$ and $\mathbb{E}[R_{\theta_{i,j}}]$. In the remainder of this section, we derive additional equations by expressing $\mathbb{E}[R_{\theta_{i,j}}]$ in terms of $\mathbb{E}[Q_{i,n}]$.

Thereto, we first focus on $\mathbb{E}[R_{\theta_{i,1}}]$. At an arbitrary moment within a station time of queue i, $Q_{i,i}$ type-i customers are waiting, who all initiate a busy period with mean $\mathbb{E}[B_i]/(1 - \rho_i)$. Furthermore, with probabilities $\rho_i\mathbb{E}[C]/\mathbb{E}[\theta_{i,1}]$ and $\mathbb{E}[S_i]/\mathbb{E}[\theta_{i,1}]$ an additional busy period with mean $\mathbb{E}[R_{B_i}]/(1 - \rho_i)$ and $\mathbb{E}[R_{S_i}]/(1 - \rho_i)$ is induced, respectively. So, we have

$$\mathbb{E}[R_{\theta_{i,1}}] = \frac{1}{1 - \rho_i}\left(\mathbb{E}[Q_{i,i}]\mathbb{E}[B_i] + \frac{\rho_i\mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{S_i}]\right), \quad i = 1, 2, \ldots, N. \quad (3.11)$$

Next, we turn our attention to $\mathbb{E}[R_{\theta_{i,2}}]$. With probability $q_{i+1,1}/q_{i,2}$, the interval $R_{\theta_{i,2}}$ is simply equal to $R_{\theta_{i+1,1}}$. On the other hand, with probability $q_{i,1}/q_{i,2}$ this residual period equals $R_{\theta_{i,1}} + S_{i+1}$ plus the busy periods initiated by the type-$(i + 1)$ customers arriving during $R_{\theta_{i,1}} + S_{i+1}$ and by the type-$(i + 1)$ customers present at an arbitrary moment within a station time of queue i. That is, we have, for $i = 1, 2, \ldots, N$,

$$\begin{aligned}\mathbb{E}[R_{\theta_{i,2}}] &= \frac{q_{i,1}}{q_{i,2}}\left((\mathbb{E}[R_{\theta_{i,1}}] + \mathbb{E}[S_{i+1}])\left(1 + \frac{\lambda_{i+1}\mathbb{E}[B_{i+1}]}{1 - \rho_{i+1}}\right) + \frac{\mathbb{E}[Q_{i+1,i}]\mathbb{E}[B_{i+1}]}{1 - \rho_{i+1}}\right) \\ &\quad + (1 - \frac{q_{i,1}}{q_{i,2}})\mathbb{E}[R_{\theta_{i+1,1}}] \\ &= \frac{q_{i,1}}{q_{i,2}}\left(\frac{\mathbb{E}[R_{\theta_{i,1}}]}{1 - \rho_{i+1}} + \frac{\mathbb{E}[S_{i+1}] + \mathbb{E}[Q_{i+1,i}]\mathbb{E}[B_{i+1}]}{1 - \rho_{i+1}}\right) + (1 - \frac{q_{i,1}}{q_{i,2}})\mathbb{E}[R_{\theta_{i+1,1}}]. \quad (3.12)\end{aligned}$$

The derivation of $\mathbb{E}[R_{\theta_{i,j}}]$ for general j proceeds along the same lines and is, therefore, omitted. Thus, we have derived the following set of MVA equations for the unknowns $\mathbb{E}[Q_{i,n}]$ and $\mathbb{E}[R_{\theta_{i,j}}]$ as summarized in the theorem below.

Theorem 3.1.1 *For $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, N-1$,*

$$\sum_{n=1}^{N} q_{n,1}\mathbb{E}[Q_{i,n}] = \frac{\lambda_i}{1-\rho_i}\Big(\rho_i\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1-q_{i,1})(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i])\Big), \tag{3.13}$$

$$\lambda_i\mathbb{E}[R_{\theta_{i+1,j}}] = \sum_{n=i+1}^{i+j}\frac{q_{n,1}}{q_{i+1,j}}\mathbb{E}[Q_{i,n}], \tag{3.14}$$

$$\mathbb{E}[R_{\theta_{i,1}}] = \frac{1}{1-\rho_i}\left(\mathbb{E}[Q_{i,i}]\mathbb{E}[B_i] + \frac{\rho_i\mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{S_i}]\right), \tag{3.15}$$

and for $j = 2, 3, \ldots, N-1$,

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{q_{i,1}}{q_{i,j}}\left(\frac{\mathbb{E}[R_{\theta_{i,1}}]}{\prod_{n=1}^{j-1}(1-\rho_{i+n})} + \sum_{n=1}^{j-1}\frac{\mathbb{E}[S_{i+n}] + \mathbb{E}[Q_{i+n,i}]\mathbb{E}[B_{i+n}]}{\prod_{m=n}^{j-1}(1-\rho_{i+m})}\right) + (1-\frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\theta_{i+1,j-1}}]. \tag{3.16}$$

□

Eliminating $\mathbb{E}[R_{\theta_{i,j}}]$ from (3.13) and (3.14) with the help of (3.15) and (3.16) renders a set of N^2 linear equations for equally many unknowns $\mathbb{E}[Q_{i,n}]$. After solving these equations, the unconditional mean queue lengths and mean delays can be computed via (3.6) and Little's Law. It is noteworthy that the residual cycle lengths $\mathbb{E}[R_{\theta_{i,N}}]$, $i = 1, 2, \ldots, N$, which are not required for the computation of the mean delays, satisfy (3.16) as well. Below we provide a significant simplification of the MVA equations by decomposing the mean waiting time into two terms, cf. [73]. The first term is a simple function of the sum of the mean setup times, whereas the second term equals the mean waiting time in a polling system obtained from the original one by modifying the service time variances and setting the mean setup times equal to zero.

Relating systems with and without setup times. We start our analysis with the following lemma, which shows the construction of an alternative system with no setup time variance, but the same conditional, and thus also unconditional, queue lengths. This result is known as the *variance absorption* result (see [187] and [73]) and its proof is remarkably simple thanks to the structure of the MVA equations.

Lemma 3.1.2 *For $i = 1, 2, \ldots, N$ and $n = 1, 2, \ldots, N$,*

$$\mathbb{E}[Q_{i,n}] = \mathbb{E}[\tilde{Q}_{i,n}], \tag{3.17}$$

where $\mathbb{E}[\tilde{Q}_{i,n}]$ is the mean conditional queue length in an exhaustive polling system with the same arrival rates, deterministic setup times with mean $\mathbb{E}[S_i]$ and service times $\tilde{B}_i$ with mean $\mathbb{E}[\tilde{B}_i] = \mathbb{E}[B_i]$ and variance $\mathbb{V}ar[\tilde{B}_i] = \mathbb{V}ar[B_i] + \frac{\mathbb{V}ar[S_i]}{\lambda_i\mathbb{E}[C]}$.

Proof. In the MVA set (3.13)-(3.16), second-order moments of both the service and setup times only show up in (3.13) and (3.15). Moreover, they only appear in the fixed combination $\rho_i\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}]$ implying that the mean conditional queue lengths remain constant as long as the value of this combination remains unchanged. It is straightforwardly verified that the alternative system defined in the lemma satisfies this condition and, thus, the proof is completed. □

In the present section, we denote the counterpart of each random variable X in the corresponding polling system with deterministic setup times by $\tilde{X}$. A direct implication of Lemma 3.1.2 is, of course, that also the mean residual station times are identical in the original and the introduced alternative system. By recalling that the term $\lambda_i\mathbb{E}[C]$ precisely equals the mean number of type-i customers served per cycle, the second term in $\mathbb{V}\mathrm{ar}[\tilde{B}_i]$ apportions the setup time variance to each of these customers fairly. Lemma 3.1.2 shows that the variance absorption result occurs at the level of the *conditional* queue lengths and that the result at the level of the *unconditional* queue lengths, as observed by [187] and [73], is a derivative. Furthermore, the proof of Lemma 3.1.2 shows that a variety of trade-offs exists between service and setup time variance, all of which yield the same conditional queue lengths and that, thus, the one stated in the main text of the lemma is just one of the possibilities.

Next, we prove the following result which states that, for *deterministic* setup times, the mean unconditional queue length can be expressed as the sum of the mean queue length in a system without setup times and a simple function of the sum of the mean setup times (see, also, [106; 198]).

Lemma 3.1.3 *For* $i = 1, 2, \ldots, N$,

$$\mathbb{E}[\tilde{Q}_i] = \mathbb{E}[\tilde{Q}_i^0] + \lambda_i\frac{\mathbb{E}[S]}{2}\frac{1-\rho_i}{1-\rho}, \tag{3.18}$$

where $\mathbb{E}[\tilde{Q}_i^0]$ *is the mean unconditional queue length in an exhaustive polling system with the same arrival rates, with zero setup times and service times* $\tilde{B}_i$.

Proof. The main ingredient of the proof is the additivity property of linear mappings. Therefore, we write the solution of the MVA set (3.13)-(3.16) as follows,

$$\begin{aligned}
\mathbb{E}[\tilde{Q}_{i,n}] &= \mathbb{E}[\tilde{Q}'_{i,n}] + \mathbb{E}[\tilde{Q}''_{i,n}], \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N, &(3.19)\\
\mathbb{E}[R_{\tilde{\theta}_{i,j}}] &= \mathbb{E}[R_{\tilde{\theta}'_{i,j}}] + \mathbb{E}[R_{\tilde{\theta}''_{i,j}}], \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N-1. &(3.20)
\end{aligned}$$

First, the unknowns $\mathbb{E}[\tilde{Q}'_{i,n}]$ and $\mathbb{E}[R_{\tilde{\theta}'_{i,j}}]$ satisfy the MVA set (3.13)-(3.16), where we set all random variables related to *setup* times equal to zero, i.e., for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N-1$,

$$\sum_{n=1}^{N} q_{n,1}\mathbb{E}[\tilde{Q}'_{i,n}] = \frac{\lambda_i}{1-\rho_i}\Big(\rho_i\mathbb{E}[R_{\tilde{B}_i}] + q_{i+1,N-1}\mathbb{E}[R_{\tilde{\theta}'_{i+1,N-1}}]\Big), \tag{3.21}$$

$$\lambda_i\mathbb{E}[R_{\tilde{\theta}'_{i+1,j}}] = \sum_{n=i+1}^{i+j}\frac{q_{n,1}}{q_{i+1,j}}\mathbb{E}[\tilde{Q}'_{i,n}], \tag{3.22}$$

$$\mathbb{E}[R_{\tilde{\theta}'_{i,1}}] = \frac{1}{1-\rho_i}\left(\mathbb{E}[\tilde{Q}'_{i,i}]\mathbb{E}[\tilde{B}_i] + \frac{\rho_i\mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{\tilde{B}_i}]\right), \tag{3.23}$$

and for $j = 2, 3, \ldots, N-1$,

$$\begin{aligned}\mathbb{E}[R_{\tilde{\theta}'_{i,j}}] &= \frac{q_{i,1}}{q_{i,j}}\left(\frac{\mathbb{E}[R_{\tilde{\theta}'_{i,1}}]}{\prod_{n=1}^{j-1}(1-\rho_{i+n})} + \sum_{n=1}^{j-1}\frac{\mathbb{E}[\tilde{Q}'_{i+n,i}]\mathbb{E}[\tilde{B}_{i+n}]}{\prod_{m=n}^{j-1}(1-\rho_{i+m})}\right) \\ &\qquad + (1-\frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\tilde{\theta}'_{i+1,j-1}}]. \qquad (3.24)\end{aligned}$$

Unfortunately, the set (3.21)-(3.24) does not allow for a closed-form solution. However, later on we derive the MVA set for systems with *zero* setup times. Foregoing the details of this derivation, we relate the solution of the set (3.21)-(3.24) to the unknowns in systems without setup times which are indicated by a superscript 0 (which is easily verified - when we present the results for systems with zero setup times - by exploiting the similarities between both sets),

$$\begin{aligned}\mathbb{E}[\tilde{Q}'_{i,i+n}] &= \frac{\rho_{i+n}}{q_{i+n,1}}\mathbb{E}[\tilde{Q}^0_{i,i+n}], && i = 1, 2, \ldots, N, \qquad (3.25)\\ & && n = 0, 1, \ldots, N-1,\\ \mathbb{E}[R_{\tilde{\theta}'_{i,j}}] &= \frac{\sum_{n=i}^{i+j-1}\rho_n}{q_{i,j}}\mathbb{E}[R_{\tilde{\theta}^0_{i,j}}], && i = 1, 2, \ldots, N, \qquad (3.26)\\ & && j = 1, 2, \ldots, N-1.\end{aligned}$$

Second, the unknowns $\mathbb{E}[\tilde{Q}''_{i,n}]$ and $\mathbb{E}[R_{\tilde{\theta}''_{i,j}}]$ satisfy the MVA set (3.13)-(3.16), where we set all residual *service* times equal to zero, i.e., for $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, N-1$,

$$\sum_{n=1}^{N} q_{n,1}\mathbb{E}[\tilde{Q}''_{i,n}] = \frac{\lambda_i}{1-\rho_i}\Big(\frac{\mathbb{E}[S_i]^2}{2\mathbb{E}[C]} + (1-q_{i,1})(\mathbb{E}[R_{\tilde{\theta}''_{i+1,N-1}}] + \mathbb{E}[S_i])\Big), \qquad (3.27)$$

$$\lambda_i\mathbb{E}[R_{\tilde{\theta}''_{i+1,j}}] = \sum_{n=i+1}^{i+j}\frac{q_{n,1}}{q_{i+1,j}}\mathbb{E}[\tilde{Q}''_{i,n}], \qquad (3.28)$$

$$\mathbb{E}[R_{\tilde{\theta}''_{i,1}}] = \frac{1}{1-\rho_i}\left(\mathbb{E}[\tilde{Q}''_{i,i}]\mathbb{E}[\tilde{B}_i] + \frac{\mathbb{E}[S_i]^2}{2\mathbb{E}[\tilde{\theta}_{i,1}]}\right), \qquad (3.29)$$

and for $j = 2, 3, \ldots, N-1$,

$$\begin{aligned}\mathbb{E}[R_{\tilde{\theta}''_{i,j}}] &= \frac{q_{i,1}}{q_{i,j}}\left(\frac{\mathbb{E}[R_{\tilde{\theta}''_{i,1}}]}{\prod_{n=1}^{j-1}(1-\rho_{i+n})} + \sum_{n=1}^{j-1}\frac{\mathbb{E}[S_{i+n}] + \mathbb{E}[\tilde{Q}''_{i+n,i}]\mathbb{E}[\tilde{B}_{i+n}]}{\prod_{m=n}^{j-1}(1-\rho_{i+m})}\right) \\ &\qquad + (1-\frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\tilde{\theta}''_{i+1,j-1}}]. \qquad (3.30)\end{aligned}$$

One can verify that the set (3.27)-(3.30) has a closed-form solution, which reads

$$\begin{aligned}\mathbb{E}[\tilde{Q}''_{i,i}] &= \frac{\lambda_i}{2}\left(\mathbb{E}[C] - \mathbb{E}[\theta_{i,1}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[C]\right), && i = 1, 2, \ldots, N, \qquad (3.31)\\ \mathbb{E}[\tilde{Q}''_{i,i+n}] &= \lambda_i\left(\mathbb{E}[\theta_{i+1,n-1}] + \frac{1}{2}\mathbb{E}[\theta_{i+n,1}]\right), && i = 1, 2, \ldots, N, \qquad (3.32)\\ & && n = 1, 2, \ldots, N-1,\\ \mathbb{E}[R_{\tilde{\theta}''_{i,j}}] &= \frac{1}{2}\mathbb{E}[\theta_{i,j}], && i = 1, 2, \ldots, N, \qquad (3.33)\\ & && j = 1, 2, \ldots, N-1,\end{aligned}$$

where $\mathbb{E}[\theta_{i+1,0}] = 0$.

It is readily seen that the additivity property implies that (3.19) and (3.20) indeed hold. Subsequently, combining the solutions (3.25)-(3.26) and (3.31)-(3.33) gives

$$\begin{aligned}\mathbb{E}[\tilde{Q}_i] &= \mathbb{E}[\tilde{Q}'_i] + \mathbb{E}[\tilde{Q}''_i] = \sum_{n=1}^{N} q_{n,1}\mathbb{E}[\tilde{Q}'_{i,n}] + \sum_{n=1}^{N} q_{n,1}\mathbb{E}[\tilde{Q}''_{i,n}] \\ &= \mathbb{E}[\tilde{Q}^0_i] + \lambda_i \frac{\mathbb{E}[S]}{2}\frac{1-\rho_i}{1-\rho}, \qquad (3.34)\end{aligned}$$

which completes the proof. □

The main consequence of the above two lemmas is that for the computation of mean waiting times in general polling systems we can focus on systems with zero setup times, for which the MVA set is significantly less intricate as shown below.

Systems without setup times. In the analysis of systems with zero setup times, we have to be careful since, if the setup times tend to zero, then each time the system becomes empty, the server will execute an infinite number of cycles in a finite time interval. This implies that the mean station times, and also the mean cycle lengths, converge to zero, which causes problems in the definition of the probabilities $q_{i,j}$. To circumvent these difficulties, we modify the definition for a mean (i,j)-period as follows

$$\mathbb{E}[\theta_{i,j}] = \sum_{n=i}^{i+j-1} \rho_n \mathbb{E}[C], \qquad i = 1,2,\ldots,N, \qquad j = 1,2,\ldots,N, \qquad (3.35)$$

where we can leave the value of $\mathbb{E}[C]$ unspecified, since it appears that, in case of zero setup times, this quantity cancels out in all steps of the analysis. Then, the probabilities $q_{i,j}$ are again well defined and change accordingly to

$$q_{i,j} = \sum_{n=i}^{i+j-1} \rho_n, \qquad i = 1,2,\ldots,N, \qquad j = 1,2,\ldots,N. \qquad (3.36)$$

By replacing all variables related to setup times by zeros and applying (3.35), the mean waiting time in polling systems with zero setup times can be computed as well. After some straightforward but tedious manipulations - omitted in the interest of brevity - which eliminate the unknowns $\mathbb{E}[R_{\theta_{i,j}}]$, the resulting MVA set looks as presented in the following theorem.

Theorem 3.1.4 *For $i = 1,2,\ldots,N$,*

$$\sum_{n=1}^{N} \rho_n \frac{1}{\lambda_i}\mathbb{E}[\tilde{Q}^0_{i,n}] - \frac{1}{\lambda_i}\mathbb{E}[\tilde{Q}^0_{i,i}] = -\mathbb{E}[R_{\tilde{B}_i}], \qquad (3.37)$$

and for $i = 1,2,\ldots,N$ and $j = 1,2,\ldots,N-1$,

$$\sum_{n=0}^{j-1} \rho_{i+1+n}\frac{1}{\lambda_i}\mathbb{E}[\tilde{Q}^0_{i,i+1+n}] + \sum_{n=0}^{j-1} \rho_{i+1+n}\frac{1}{\lambda_{i+j}}\mathbb{E}[\tilde{Q}^0_{i+j,i+1+n}] - \frac{1}{\lambda_i}\mathbb{E}[\tilde{Q}^0_{i,i+j}] = -\mathbb{E}[R_{\tilde{B}_{i+j}}]. \qquad (3.38)$$

□

The unconditional mean queue lengths can, subsequently, be obtained from

$$\mathbb{E}[\tilde{Q}_i^0] = \sum_{n=1}^{N} \rho_n \mathbb{E}[\tilde{Q}_{i,n}^0], \qquad i = 1, 2, \ldots, N. \tag{3.39}$$

Now, (3.37) and (3.38) form a set of N^2 linear equations for equally many unknowns $\mathbb{E}[\tilde{Q}_{i,n}^0]$. After solving these equations, the unconditional mean queue lengths and mean waiting times can be computed via (3.39) and Little's Law. Application of Lemmas 3.1.2 and 3.1.3, subsequently, yields the mean queue lengths and waiting times for the corresponding polling system with setup times (possibly for a number of setup time scenarios).

It is important to remark that the coefficient matrix of the MVA set formed by (3.37) and (3.38) has N^4 elements of which only a fraction $1/N$ is nonzero implying that this matrix is very sparse, which can be exploited in the efficient storage and manipulation of this set. Below we illustrate the above procedure in the evaluation of a specific system, but first we present a remark.

Remark 3.1.5 In the case of a *symmetric* system, such that all queues are statistically identical, (3.37) and (3.38) allow for a closed-form solution of the mean conditional queue lengths, i.e., we have for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N-1$,

$$\mathbb{E}[\tilde{Q}_{i,i}^0] = \frac{\lambda_i}{1-\rho}\mathbb{E}[R_{\tilde{B}_i}], \quad \text{and} \quad \mathbb{E}[\tilde{Q}_{i,i+j}^0] = \frac{\lambda_i(1-(N+1-2j)\rho_i)}{(1-\rho)(1-\rho_i)}\mathbb{E}[R_{\tilde{B}_i}], \tag{3.40}$$

which can be verified by substitution. □

Example 3.1.6 Let us consider a general two-queue exhaustive polling system without setup times. The set of equations formed by (3.37) and (3.38) can be written in matrix form as

$$\begin{pmatrix} 1-\rho_1 & -\rho_2 & 0 & 0 \\ 0 & 1-\rho_2 & 0 & -\rho_2 \\ -\rho_1 & 0 & 1-\rho_1 & 0 \\ 0 & 0 & -\rho_1 & 1-\rho_2 \end{pmatrix} \begin{pmatrix} \frac{1}{\lambda_1}\mathbb{E}[\tilde{Q}_{1,1}^0] \\ \frac{1}{\lambda_1}\mathbb{E}[\tilde{Q}_{1,2}^0] \\ \frac{1}{\lambda_2}\mathbb{E}[\tilde{Q}_{2,1}^0] \\ \frac{1}{\lambda_2}\mathbb{E}[\tilde{Q}_{2,2}^0] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[R_{\tilde{B}_1}] \\ \mathbb{E}[R_{\tilde{B}_2}] \\ \mathbb{E}[R_{\tilde{B}_1}] \\ \mathbb{E}[R_{\tilde{B}_2}] \end{pmatrix}. \tag{3.41}$$

The coefficient matrix $\mathbf{A}$ on the lefthand side possesses the following determinant,

$$\det(\mathbf{A}) = (1-\rho)(1-\rho+2\rho_1\rho_2), \tag{3.42}$$

which clearly indicates that the MVA equations have a unique solution if $\rho < 1$ (i.e., when the system is stable). □

3.1.3 Pseudo-conservation law

We now return to the general exhaustive polling system with stochastic setup times and aim to derive the pseudo-conservation law in such a system. First of all, Lemmas 3.1.2 and 3.1.3 give the following expression for the mean total amount of work in the original system with stochastic setup times,

$$\sum_{i=1}^{N} \mathbb{E}[B_i]\mathbb{E}[Q_i] = \sum_{i=1}^{N} \mathbb{E}[B_i]\mathbb{E}[\tilde{Q}_i^0] + + \frac{\mathbb{E}[S]}{2(1-\rho)} \sum_{i=1}^{N} \rho_i(1-\rho_i). \tag{3.43}$$

The first term in the righthand side of (3.43) equals the mean total amount of work in the transformed system without setup times (excluding the customer possibly in service). Due to the work-conservative nature of the exhaustive discipline, this amount equals the corresponding amount of work in an FCFS $M/G/1$, in which all customer classes are lumped together into one customer class with arrival rate $\Lambda = \lambda_1 + \lambda_2 + \dots \lambda_N$ and service times with mean $\sum_{i=1}^N \frac{\lambda_i}{\Lambda}\mathbb{E}[\tilde{B}_i]$ and second moment $\sum_{i=1}^N \frac{\lambda_i}{\Lambda}\mathbb{E}[\tilde{B}_i^2]$. That is,

$$\sum_{i=1}^N \mathbb{E}[B_i]\mathbb{E}[\tilde{Q}_i^0] = \rho \sum_{i=1}^N \frac{\lambda_i \mathbb{E}[\tilde{B}_i^2]}{2(1-\rho)} = \rho \sum_{i=1}^N \frac{\lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)} + \frac{\rho}{2\mathbb{E}[S]}\mathbb{V}\text{ar}[S]. \tag{3.44}$$

Since the last term in the righthand side of (3.43) only depends on known input parameters, this leads to the pseudo-conservation law for exhaustive polling systems (after application of Little's Law),

$$\sum_{i=1}^N \lambda_i \mathbb{E}[W_i] = \rho \sum_{i=1}^N \frac{\lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)} + \frac{\rho}{2\mathbb{E}[S]}\mathbb{V}\text{ar}[S] + \frac{\mathbb{E}[S]}{2(1-\rho)}\sum_{i=1}^N \rho_i(1-\rho_i). \tag{3.45}$$

For more information on pseudo-conservation laws, we refer to Chapter 2.

3.1.4 Correlations

The set of MVA equations can also be applied for the computation of the mean of residual (i,j)-periods $\mathbb{E}[R_{\theta_{i,j}}]$. Thereupon, the variance of an (i,j)-period can be obtained via

$$\mathbb{V}\text{ar}[\theta_{i,j}] = 2\mathbb{E}[R_{\theta_{i,j}}]\mathbb{E}[\theta_{i,j}] - \mathbb{E}[\theta_{i,j}]^2, \qquad i = 1,2,\dots,N, \quad j = 1,2,\dots,N, \tag{3.46}$$

where $\mathbb{E}[\theta_{i,j}]$ is given by (3.3). From these variances it is also possible to compute the covariance $\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}]$ and correlation $\mathbb{C}\text{or}[\theta_i, \theta_{i+n}]$ of the station periods θ_i and θ_{i+n} via the following lemma.

Lemma 3.1.7 *Given random variables X, Y and Z, the covariance of X and Z can be obtained as follows*

$$\mathbb{C}ov[X,Z] = \frac{1}{2}(\mathbb{V}ar[X+Y+Z] - \mathbb{V}ar[X+Y] - \mathbb{V}ar[Y+Z] + \mathbb{V}ar[Y]).$$

Proof. By definition,

$$\mathbb{V}\text{ar}[X+Y+Z] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y] + \mathbb{V}\text{ar}[Z] + 2(\mathbb{C}\text{ov}[X,Y] + \mathbb{C}\text{ov}[Y,Z] + \mathbb{C}\text{ov}[X,Z]),$$

and

$$\begin{aligned} \mathbb{V}\text{ar}[X+Y] &= \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y] + 2\mathbb{C}\text{ov}[X,Y], \\ \mathbb{V}\text{ar}[Y+Z] &= \mathbb{V}\text{ar}[Y] + \mathbb{V}\text{ar}[Z] + 2\mathbb{C}\text{ov}[Y,Z]. \end{aligned}$$

which, after some rewriting, completes the proof. □

Using Lemma 3.1.7, one may verify that for $i = 1,2,\dots,N$ and $n = 1,2,\dots,N-1$,

$$\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}] = \frac{1}{2}(\mathbb{V}\text{ar}[\theta_{i,n+1}] - \mathbb{V}\text{ar}[\theta_{i,n}] - \mathbb{V}\text{ar}[\theta_{i+1,n}] + \mathbb{V}\text{ar}[\theta_{i+1,n-1}]), \tag{3.47}$$

where $\mathbb{V}\text{ar}[\theta_{i+1,0}] = 0$. Finally, the correlation follows by using

$$\mathbb{C}\text{or}[\theta_i, \theta_{i+n}] = \frac{\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}]}{\sqrt{\mathbb{V}\text{ar}[\theta_i]\mathbb{V}\text{ar}[\theta_{i+n}]}}, \qquad i = 1,2,\dots,N, \quad n = 1,2,\dots,N-1. \tag{3.48}$$

Following the above analysis, we can prove that in a *symmetric* system with setup times these correlations admit the following remarkably simple form for $i = 1, 2, \ldots, N$ and $n = 1, 2, \ldots, N-1$,

$$\mathbb{C}\text{or}[\theta_i, \theta_{i+n}] = \frac{\rho}{N - (N-1)\rho}. \tag{3.49}$$

It immediately strikes the eye that the correlations between station times remain constant (within a single cycle) and that they are independent of the setup times and the service time distributions. Furthermore, they are always positive and monotone increasing (decreasing) in the total load ρ (in the number of queues N).

3.1.5 Model variations

The present subsection shows that MVA can directly be generalized to a variety of model variations of the exhaustive system analyzed in Subsection 3.1.2, i.e., systems with gated service, systems with mixed service and systems with fully gated service (see [38]).

Gated service. In case of gated service, all customers waiting in queue at the start of a station time of this queue are placed behind a gate meaning that they are served in the current cycle. However, customers arriving during a station time of their queue are placed before this gate and are, thus, only served in the next cycle. With this difference understood, it is clear that, in case $i = n$, $Q_{i,n}$ is the sum of two auxiliary variables, i.e.,

$$Q_{i,i} = \bar{Q}_{i,i} + \hat{Q}_{i,i}, \qquad i = 1, 2, \ldots, N, \tag{3.50}$$

where $\bar{Q}_{i,i}$ and $\hat{Q}_{i,i}$ represent the queue length behind and before the gate, respectively. Recall that the customer in service is excluded. In case $i \neq n$, all customers in queue i are obviously located before the gate, i.e.,

$$Q_{i,n} = \hat{Q}_{i,n}, \qquad i \neq n = 1, 2, \ldots, N. \tag{3.51}$$

The corresponding unconditional queue length Q_i has mean

$$\mathbb{E}[Q_i] = \mathbb{E}[\hat{Q}_i] + q_{i,1}\mathbb{E}[\bar{Q}_{i,i}] = \sum_{n=1}^{N} q_{n,1}\mathbb{E}[\hat{Q}_{i,n}] + q_{i,1}\mathbb{E}[\bar{Q}_{i,i}], \qquad i = 1, 2, \ldots, N. \tag{3.52}$$

Again, our analysis extensively makes use of Little's Law and the PASTA property. We tag a customer at its arrival to queue i, $i = 1, 2, \ldots, N$. By the PASTA property, we know that this customer sees the system in equilibrium. So, the tagged customer has to wait for the service of all customers $\hat{Q}_i$, who were already waiting before the gate on his arrival. Furthermore, he has to wait until the first polling instant of queue i equalling a residual (i, N)-period, i.e., a residual cycle. By definition of the gated policy, this extra waiting time is incurred even in case the tagged type-i customer arrives in a station time of queue i. Consequently, the mean waiting time $\mathbb{E}[W_i]$ of a type-i customer is given by

$$\mathbb{E}[W_i] = \mathbb{E}[\hat{Q}_i]\mathbb{E}[B_i] + \mathbb{E}[R_{\theta_{i,N}}], \quad i = 1, 2, \ldots, N, \tag{3.53}$$

which, in combination with Little's Law gives us the following relation

$$\mathbb{E}[Q_i] = \rho_i\mathbb{E}[\hat{Q}_i] + \lambda_i\mathbb{E}[R_{\theta_{i,N}}], \quad i = 1, 2, \ldots, N. \tag{3.54}$$

Once more, the mean residual periods have to be obtained, where we choose the same solution approach as in the exhaustive case.

The gated policy, together with the definition of a station time, clearly implies that the number of type-i customers before the gate at an arbitrary moment within an (i,j)-period is equal to the number of Poisson arrivals during the age of an (i,j)-period, which is in distribution again equal to a residual (i,j)-period. That is,

$$\sum_{n=i}^{i+j-1} \frac{q_{n,1}}{q_{i,j}} \mathbb{E}[\hat{Q}_{i,n}] = \lambda_i \mathbb{E}[R_{\theta_{i,j}}], \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N. \tag{3.55}$$

Secondly, if we substitute (3.52) into (3.54), we get

$$(1-\rho_i) \sum_{n=1}^{N} q_{n,1} \mathbb{E}[\hat{Q}_{i,n}] + q_{i,1} \mathbb{E}[\bar{Q}_{i,i}] = \lambda_i \mathbb{E}[R_{\theta_{i,N}}], \qquad i = 1, 2, \ldots, N. \tag{3.56}$$

Now, we see that (3.55) and (3.56) comprise a set of $N(N+1)$ linear equations for $\mathbb{E}[\bar{Q}_{i,i}]$, $\mathbb{E}[\hat{Q}_{i,n}]$ and $\mathbb{E}[R_{\theta_{i,j}}]$. To eliminate the unknown mean residual (i,j)-periods from this set, below these quantities are rewritten in terms of $\mathbb{E}[\bar{Q}_{i,i}]$ and $\mathbb{E}[\hat{Q}_{i,n}]$.

Starting with $\mathbb{E}[R_{\theta_{i,1}}]$, we recognize that this period lasts at least the sum of the service times of the customers behind the gate. With probability $\rho_i \mathbb{E}[C]/\mathbb{E}[\theta_{i,1}]$ a residual service time and a setup time for queue $i+1$ is induced, while with probability $\mathbb{E}[S_{i+1}]/\mathbb{E}[\theta_{i,1}]$ only a residual setup time for queue $i+1$ is generated. Consequently, we have

$$\mathbb{E}[R_{\theta_{i,1}}] = \mathbb{E}[\bar{Q}_{i,i}]\mathbb{E}[B_i] + \frac{\mathbb{E}[S_{i+1}]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{S_{i+1}}] + \frac{\rho_i \mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]}(\mathbb{E}[R_{B_i}] + \mathbb{E}[S_{i+1}]), \quad i = 1, 2, \ldots, N. \tag{3.57}$$

In case of an $(i,2)$-period, $R_{\theta_{i,2}}$ equals $R_{\theta_{i+1,1}}$ with probability $q_{i+1,1}/q_{i,2}$. With probability $q_{i,1}/q_{i,2}$, however, this residual period equals $R_{\theta_{i,1}} + S_{i+2}$ plus the service times of the type-$(i+1)$ customers present at an arbitrary moment within a station time of queue i and of the type-$(i+1)$ customers arriving during $R_{\theta_{i,1}}$. This yields, for $i = 1, 2, \ldots, N$,

$$\begin{aligned} \mathbb{E}[R_{\theta_{i,2}}] &= \frac{q_{i,1}}{q_{i,2}} \Big(\mathbb{E}[R_{\theta_{i,1}}] + \mathbb{E}[S_{i+2}] + (\lambda_{i+1}\mathbb{E}[R_{\theta_{i,1}}] + \mathbb{E}[\hat{Q}_{i+1,i}])\mathbb{E}[B_{i+1}] \Big) + \\ &\qquad\qquad (1 - \frac{q_{i,1}}{q_{i,2}})\mathbb{E}[R_{\theta_{i+1,1}}] \\ &= \frac{q_{i,1}}{q_{i,2}} \Big(\mathbb{E}[R_{\theta_{i,1}}](1+\rho_{i+1}) + \mathbb{E}[S_{i+2}] + \mathbb{E}[\hat{Q}_{i+1,i}]\mathbb{E}[B_{i+1}] \Big) + \\ &\qquad\qquad (1 - \frac{q_{i,1}}{q_{i,2}})\mathbb{E}[R_{\theta_{i+1,1}}]. \end{aligned} \tag{3.58}$$

The derivation of $\mathbb{E}[R_{\theta_{i,j}}]$ for general j is similar. After some straightforward calculations, the following expression can be derived for $i = 1, 2, \ldots, N$ and $j = 2, 3, \ldots, N$,

$$\begin{aligned} \mathbb{E}[R_{\theta_{i,j}}] &= \frac{q_{i,1}}{q_{i,j}} \Big(\mathbb{E}[R_{\theta_{i,1}}] \prod_{n=1}^{j-1}(1+\rho_{i+n}) + \sum_{n=1}^{j-1}(\mathbb{E}[S_{i+n+1}] + \\ &\qquad \mathbb{E}[\hat{Q}_{i+n,i}]\mathbb{E}[B_{i+n}]) \prod_{m=n+1}^{j-1}(1+\rho_{i+m}) \Big) + (1 - \frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\theta_{i+1,j-1}}]. \end{aligned} \tag{3.59}$$

To conclude, elimination of $\mathbb{E}[R_{\theta_{i,j}}]$ from (3.55) and (3.56) with the help of (3.57) and (3.59) yields a set of $N(N+1)$ linear equations for equally many unknowns $\mathbb{E}[\bar{Q}_{i,i}]$ and $\mathbb{E}[\hat{Q}_{i,n}]$. Together with (3.52) and Little's Law, the solution to these equations yields the unconditional mean queue lengths and mean waiting times.

Systems with mixed service. To treat the case of mixed service polling systems, we first have to realize that the two different definitions of the station times conflict. It is, therefore, necessary to do the conditioning of the queue lengths on the system state in a more detailed manner. That is, we introduce $M_{i,n}$ and $N_{i,n}$ as the queue lengths at queue i at an arbitrary epoch within a setup time and a service period of queue n, $i, n = 1, 2, \ldots, N$, respectively. Recall that in the purely exhaustive and gated systems, we could aggregate a setup time and a service period in a single random variable, the station period. It goes without saying that for the gated queues we again have to distinguish between the queue length behind and before the gate. The total number of variables, and thus the total number of linear equations, is now equal to $2N^2 + K$ with $K \leq N$ the number of queues deploying the gated discipline.

Fully gated strategy. A minor but interesting variant of the gated discipline is the so-called fully gated strategy [38], also called the reserved gated strategy [47]. Under this fully gated policy, all, and only, customers found by the server at the start of the setup time are served. For this fully gated strategy, the definition of the station time can be chosen identical to the definition for exhaustive service, which simplifies the analysis of a mixed exhaustive/fully gated system considerably (the conditioning of the queue lengths on the system state can obviously be done in the standard manner of Subsection 3.1.1).

Finally, we want to stress that for all of the above model variations (and, actually, many more as touched upon in Section 3.5) similar results as in the exhaustive case, such as pseudo-conservation laws, could be derived. In a similar vein, the remainder of the present chapter focusses on the exhaustive discipline, but all results can straightforwardly be applied to the above model variations. We omit, however, these results in the interest of space and refer the interested reader to the corresponding papers [P4; P5; P9; P10; P15; P16; P17].

3.2 Increasing load

The present section is a shortened version of [P5], of which some preliminary results appeared in [P4]. We consider the exact asymptotic analysis of the mean waiting time in exhaustive polling systems in heavy traffic, i.e., as ρ tends to one. In the past several approaches have been suggested for deriving heavy-traffic asymptotics in polling systems. Coffman *et al.* [66; 67] use a heavy-traffic averaging principle to study a two-queue model with exhaustive service at both queues and show that, under heavy-traffic assumptions and scalings, the total amount of unfinished work converges to a known process. These observations lead to explicit expressions for the moments of the waiting time at both queues. They also suggest that, based on a partial conjecture, the analysis can be extended to systems with more than two queues. Exploring the averaging principle, Reiman and Wein [181] and Markowitz *et al.* [157; 158] study the problem of determining optimal dynamic schedules by approximating the dynamic scheduling problems by diffusion control problems (see, also, the literature review on the SELSP in Chapter 2). Kudoh *et al.* [137] use the classical buffer-occupancy technique, which is based on an expression for the probability generating function of the joint queue length distribution at successive polling instants, to derive explicit expressions for the second moment of the waiting time in fully symmetric systems with gated or exhaustive service at each queue for models with two, three and four queues. They also give conjectures about the heavy-traffic limits of the first two moments of the waiting time for systems with an arbitrary number of queues. Kroese [136] uses the theory of age-dependent branching processes to study the heavy-traffic behavior of continuous polling systems and shows that the steady-state number of waiting customers has approximately a gamma distribution. Van der Mei and co-authors [161; 162; 164; 167;

174; 175] explore the recursive relations of the descendant set approach to derive closed-form expressions for the asymptotic waiting time distribution in heavy traffic for polling systems with a multi-type branching structure both for cyclic and periodic server routing. Most results in these papers are generalized and unified in [165] (see, also, Chapter 2).

We propose a new technique to derive heavy-traffic limits for the expected waiting time at each of the queues by using the MVA analysis of the preceding section as the starting point. By taking the proper heavy-traffic limits of the MVA set, we obtain a highly simplified but dependent set of linear equations that determines the heavy-traffic limits of $\mathbb{E}[Q_{i,n}]$ up to a scaling constant. Finally, the scaling constant is obtained by adding a linear equation that follows from the pseudo-conservation law of the system cf., (3.45). These results do not only provide a new means to obtain heavy-traffic asymptotics for the expected waiting time, but also lead to the observation that the correlations between successive station times converge to one as the load tends to one. The latter observation gives rise to asymptotic expressions for the covariance between successive station times. The results of the present section could be used as approximate closed-form expressions for stable systems, i.e., with load less than one, allowing for back-of-the-envelope calculations. Numerical results in [P4; P5] show that such approximations are accurate when the load is roughly 90% or more.

The remainder of this section is organized as follows. Section 3.2.1 presents some additional notation. Subsections 3.2.2 and 3.2.3 analyze the MVA equations in heavy traffic and obtain closed-form expressions for the mean asymptotic waiting time and the covariances between successive station times, respectively. Finally, in Subsection 3.2.4 we show that our results lead to new (managerial) insights in the impact of high load on exhaustive lot-sizing policies.

3.2.1 Notation

Throughout the present section the mean waiting time $\mathbb{E}[W_i]$ of a type-i customer, $i = 1, 2, \dots, N$, is considered as a function of ρ, where the arrival rates are variable, while the service time distributions and the ratios of the arrival rates are fixed. In case $\rho \uparrow 1$, all queues become unstable and, thus, $\mathbb{E}[W_i]$ tends to infinity for all i. To be precise, $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$ (cf. [P6; P7]),

$$\mathbb{E}[W_i] = \frac{\mathbb{E}[W_i^*]}{1-\rho} + o((1-\rho)^{-1}), \qquad \rho \uparrow 1, \quad i = 1, 2, \dots, N, \tag{3.60}$$

where $g(x) = o(f(x))$ means that $g(x)/f(x) \to 0$ as $x \uparrow 1$. The analysis of the present section is oriented towards the determination of a closed-form expression for

$$\mathbb{E}[W_i^*] = \lim_{\rho \uparrow 1} (1-\rho)\mathbb{E}[W_i], \qquad i = 1, 2, \dots, N, \tag{3.61}$$

which is referred to as the *mean asymptotic scaled waiting time* at queue i. More colloquially, we can say that $\mathbb{E}[W_i^*]$ indicates the rate at which $\mathbb{E}[W_i]$ tends to infinity as $\rho \uparrow 1$.

The fact that $\mathbb{E}[W_i]$, $i = 1, 2, \dots, N$, has a first-order pole at $\rho = 1$ implies that $\mathbb{E}[Q_{i,n}]$, $i, n = 1, 2, \dots, N$, has a first-order pole at $\rho = 1$ as well. Therefore, the following limits are well defined,

$$\mathbb{E}[Q_{i,n}^*] = \lim_{\rho \uparrow 1} (1-\rho)\mathbb{E}[Q_{i,n}], \qquad i = 1, 2, \dots, N, \quad n = 1, 2, \dots, N. \tag{3.62}$$

Finally, for each variable x that is a function of ρ, its value evaluated at $\rho = 1$ is denoted by $\hat{x}$.

3.2.2 Mean value analysis

In the present subsection we explore the use of the MVA framework to derive heavy-traffic asymptotics for the model with exhaustive service. Starting point is the MVA set given by (3.13)-(3.16), which in general can not be solved in closed-form, but below an explicit solution is derived in the limit of $\rho \uparrow 1$. Multiplying both sides of (3.13) - (3.16) by $(1-\rho)$ and letting $\rho \uparrow 1$ renders the corresponding set of equations in heavy traffic for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, N-1$,

$$\begin{aligned}
\sum_{n=1}^{N} \hat{\rho}_n \mathbb{E}[Q^*_{i,n}] &= \hat{\lambda}_i \mathbb{E}[R_{\theta^*_{i+1,N-1}}], && (3.63)\\
\frac{\sum_{n=i+1}^{i+j} \hat{\rho}_n \mathbb{E}[Q^*_{i,n}]}{\sum_{m=i+1}^{i+j} \hat{\rho}_m} &= \hat{\lambda}_i \mathbb{E}[R_{\theta^*_{i+1,j}}], && (3.64)\\
\mathbb{E}[R_{\theta^*_{i,1}}] &= \frac{1}{1-\hat{\rho}_i} \mathbb{E}[Q^*_{i,i}]\mathbb{E}[B_i], && (3.65)
\end{aligned}$$

and for $j = 2, 3, \dots, N$,

$$\begin{aligned}
\mathbb{E}[R_{\theta^*_{i,j}}] = {} & \frac{\hat{\rho}_i}{\sum_{n=i}^{i+j-1} \hat{\rho}_n} \left(\frac{\mathbb{E}[R_{\theta^*_{i,1}}]}{\prod_{n=1}^{j-1}(1-\hat{\rho}_{i+n})} + \sum_{n=1}^{j-1} \frac{\mathbb{E}[Q^*_{i+n,i}]\mathbb{E}[B_{i+n}]}{\prod_{m=n}^{j-1}(1-\hat{\rho}_{i+m})} \right) \\
& + \frac{\sum_{n=i+1}^{i+j-1} \hat{\rho}_n}{\sum_{n=i}^{i+j-1} \hat{\rho}_n} \mathbb{E}[R_{\theta^*_{i+1,j-1}}]. \quad (3.66)
\end{aligned}$$

The variables $\mathbb{E}[R_{\theta^*_{i,j}}]$ are defined by

$$\mathbb{E}[R_{\theta^*_{i,j}}] = \lim_{\rho \uparrow 1} (1-\rho)\mathbb{E}[R_{\theta_{i,j}}], \qquad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N. \qquad (3.67)$$

The fact that $\mathbb{E}[W_i]$, $i = 1, 2, \dots, N$, has a first-order pole at $\rho = 1$ implies that $\mathbb{E}[R_{\theta_{i,j}}]$, $j = 1, 2, \dots, N$, also has a first-order pole at $\rho = 1$ and, thus, the limits in (3.67) are well defined.

After some matrix manipulations, the following set of N^2 equations for equally many unknowns $\mathbb{E}[Q^*_{i,n}]$ is obtained, for $i = 1, 2, \dots, N$,

$$\sum_{n=1}^{N} \hat{\rho}_n \mathbb{E}[Q^*_{i,n}] - \mathbb{E}[Q^*_{i,i}] = 0, \qquad (3.68)$$

and for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, N-1$,

$$\sum_{n=0}^{j-1} \hat{\rho}_{i+1+n} \frac{1}{\lambda_i} \mathbb{E}[Q^*_{i,i+1+n}] + \sum_{n=0}^{j-1} \hat{\rho}_{i+1+n} \frac{1}{\lambda_{i+j}} \mathbb{E}[Q^*_{i+j,i+1+n}] - \frac{1}{\lambda_i} \mathbb{E}[Q^*_{i,i+j}] = 0, \qquad (3.69)$$

The set (3.68) - (3.69) can be solved up to some unknown scaling factor $c \in \mathbb{R}$ as shown in the following theorem.

Theorem 3.2.1 *The solution of the set* (3.68) - (3.69) *is given by*

$$\begin{aligned}
\mathbb{E}[Q^*_{i,i}] &= c\hat{\lambda}_i(1-\hat{\rho}_i), && i = 1, 2, \dots, N, && (3.70)\\
\mathbb{E}[Q^*_{i,i+n}] &= c\hat{\lambda}_i(2\sum_{m=1}^{n-1} \hat{\rho}_{i+m} + \hat{\rho}_{i+n}), && i = 1, 2, \dots, N, \quad n = 1, 2, \dots, N-1, && (3.71)
\end{aligned}$$

with $c \in \mathbb{R}$.

Proof. By Cramer's rule we know that the *homogeneous* set (3.68) - (3.69) has (an infinite number of) non-degenerate solutions if and only if the determinant of the coefficient matrix vanishes. Substitution shows that (3.70) and (3.71) is indeed a solution of this set. After elementary, but tedious, row and column operations, the final row reduced form of the coefficient matrix shows that the rank of the matrix equals $N^2 - 1$, which completes the proof. □

Since the dimension of the null space of the coefficient matrix of (3.68) - (3.69) equals one, adding the *non-homogeneous* pseudo-conservation law (3.45) gives a unique solution for the unknown scaling factor c as done in the theorem below.

Theorem 3.2.2 *The quantity c is given by*

$$c = \frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta}, \tag{3.72}$$

where $\beta = \frac{\mathbb{E}[B]}{\mathbb{E}[B^2]}$ and $\delta = 1 - \sum_{i=1}^{N} \hat{\rho}_i^2$.

Proof. Combining Theorem 3.2.1, (3.6) and Little's Law yields the unconditional mean asymptotic scaled waiting times,

$$\mathbb{E}[W_i^*] = c(1 - \hat{\rho}_i), \qquad i = 1, 2, \ldots, N, \tag{3.73}$$

which satisfy a scaled version of pseudo-conservation law (3.45). Thus, multiplying both sides of (3.45) by $(1 - \rho)$ and letting $\rho \uparrow 1$ yields

$$\sum_{i=1}^{N} \hat{\rho}_i \mathbb{E}[W_i^*] = \frac{1 + \beta\delta\mathbb{E}[S]}{2\beta}, \tag{3.74}$$

where $\beta = \frac{\mathbb{E}[B]}{\mathbb{E}[B^2]}$ and $\delta = 1 - \sum_{i=1}^{N} \hat{\rho}_i^2$. Combining (3.73) and (3.74) completes the proof. □

Theorem 3.2.2 together with (3.73) brings us in the position to obtain a closed-form expression for the mean asymptotic scaled waiting time $\mathbb{E}[W_i^*]$ at each of the queues as exposed in the following corollary.

Corollary 3.2.3 *For $i = 1, 2, \ldots, N$,*

$$\mathbb{E}[W_i^*] = (1 - \hat{\rho}_i)\frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta}. \tag{3.75}$$

Corollary 3.2.3 is in agreement with results of [67] and [166] and explicitly reveals the impact of the system parameters on the mean asymptotic scaled waiting time as stated in the following property (see also [167]).

Property 3.2.4 *For $i = 1, 2, \ldots, N$,*

1. *$\mathbb{E}[W_i^*]$ is independent of the visit order;*

2. *$\mathbb{E}[W_i^*]$ depends on the service time distributions only through the first two moments of the service time of an arbitrary customer;*

3. *$\mathbb{E}[W_i^*]$ depends on the setup time distributions only through the first moment of the total setup time in a cycle.*

□

It is important to note that the properties discussed above are in general not valid for stable systems, i.e., with $\rho < 1$.

3.2.3 Correlations

Similar to the mean waiting time $\mathbb{E}[W_i]$, we study the correlation $\mathbb{C}\text{or}[\theta_i, \theta_{i+n}]$ of the station periods θ_i and θ_{i+n} as a function of ρ, where the arrival rates are variable. Observe that (3.65) and (3.66), together with the results for the mean conditional queue lengths, yield the following expression for the mean asymptotic scaled residual (i,j)-period $\mathbb{E}[R_{\theta^*_{i,j}}]$,

$$\mathbb{E}[R_{\theta^*_{i,j}}] = \frac{1+\beta\delta\mathbb{E}[S]}{2\beta\delta} \sum_{m=i}^{i+j-1} \hat{\rho}_m, \qquad i = 1,2,\ldots,N, \quad j = 1,2,\ldots,N. \tag{3.76}$$

Combining (3.46) and (3.76) together with the following obvious observation for the mean of a scaled asymptotic (i,j)-period $\mathbb{E}[\theta^*_{i,j}]$,

$$\mathbb{E}[\theta^*_{i,j}] = \lim_{\rho\uparrow 1} (1-\rho)\mathbb{E}[\theta_{i,j}] = \mathbb{E}[S]\left(\sum_{m=i}^{i+j-1} \hat{\rho}_m\right), \qquad i = 1,2,\ldots,N, \quad j = 1,2,\ldots,N, \tag{3.77}$$

yields for the corresponding scaled asymptotic variance $\mathbb{V}\text{ar}[\theta^*_{i,j}]$,

$$\mathbb{V}\text{ar}[\theta^*_{i,j}] = \lim_{\rho\uparrow 1} (1-\rho)^2\mathbb{V}\text{ar}[\theta_{i,j}] = \frac{\mathbb{E}[S]}{\beta\delta}\left(\sum_{m=i}^{i+j-1} \hat{\rho}_m\right)^2, \qquad i = 1,2,\ldots,N, \quad j = 1,2,\ldots,N. \tag{3.78}$$

The above expression in conjunction with (3.47) and (3.48) gives rise to the following result for the scaled asymptotic covariance $\mathbb{C}\text{ov}[\theta^*_i, \theta^*_{i+n}]$ and the asymptotic correlation $\mathbb{C}\text{or}[\theta^*_i, \theta^*_{i+n}]$ of the station periods θ_i and θ_{i+n} under heavy traffic defined by, respectively,

$$\begin{aligned}
\mathbb{C}\text{ov}[\theta^*_i, \theta^*_{i+n}] &= \lim_{\rho\uparrow 1} (1-\rho)^2\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}], \ i = 1,2,\ldots,N, \ n = 1,2,\ldots,N-1, & (3.79)\\
\mathbb{C}\text{or}[\theta^*_i, \theta^*_{i+n}] &= \lim_{\rho\uparrow 1} \mathbb{C}\text{or}[\theta_i, \theta_{i+n}], \qquad i = 1,2,\ldots,N, \quad n = 1,2,\ldots,N-1, & (3.80)
\end{aligned}$$

with the remark that these limits are again well defined due to the fact that $\mathbb{E}[W_i]$, $i = 1,2,\ldots,N$, has a first-order pole at $\rho = 1$.

Corollary 3.2.5 *For $i = 1,2,\ldots,N$ and $n = 1,2,\ldots,N-1$, we have*

$$\begin{aligned}
\mathbb{C}ov[\theta^*_i, \theta^*_{i+n}] &= \frac{\hat{\rho}_i\hat{\rho}_{i+n}}{\beta\delta}\mathbb{E}[S], & (3.81)\\
\mathbb{C}or[\theta^*_i, \theta^*_{i+n}] &= 1. & (3.82)
\end{aligned}$$

□

From Corollary 3.2.5 the following properties about the dependence of the scaled asymptotic covariance and the asymptotic correlation with respect to the system parameters can be perceived.

Property 3.2.6 *For $i = 1,2,\ldots,N$ and $n = 1,2,\ldots,N-1$,*

1. *$\mathbb{C}ov[\theta^*_i, \theta^*_{i+n}]$ and $\mathbb{C}or[\theta^*_i, \theta^*_{i+n}]$ are independent of the visit order;*

2. *the station time θ_i of queue i is perfectly correlated with the station time θ_{i+n} of queue $i+n$;*

3. *$\mathbb{C}ov[\theta_i^*, \theta_{i+n}^*]$ depends on the service time distributions only through the first two moments of the service time of an arbitrary customer;*

4. *$\mathbb{C}ov[\theta_i^*, \theta_{i+n}^*]$ depends on the setup time distributions only through the first moment of the total setup time in a cycle.*

□

Notice that for *symmetric* systems, Property 3.2.6 directly follows from (3.49). Finally, for heavy-traffic results for gated service obtained via the MVA framework we refer to [P4; P5]. In these papers, also closed-form approximations - based on the heavy-traffic asymptotics - for various performance measures have been suggested and numerically tested. We close this subsection with some remarks.

Remark 3.2.7 Corollary 3.2.5 shows that the scaled asymptotic covariance of successive station times equals zero in systems with zero setup times. This observation actually holds for the covariances of station times in stable systems as well. That is, in systems without setup times the number of visits with zero length tends to infinity and, consequently, the mean and variance of (i, j)-periods both tend to zero implying that the covariance of successive station periods converges to zero. □

Remark 3.2.8 Corollary 3.2.5 can intuitively be explained from the results of Coffman *et al.* [66; 67]. They prove a heavy-traffic averaging principle (HTAP) for a two-queue polling system with exhaustive service at both queues, from which they conjecture that the same result applies for systems with more than two queues. This HTAP says that, in heavy traffic, the total workload in the system converges to a known process, while on the time scale of this process, the individual workloads change at an infinite rate. This means that the work is shifting between the queues in a rather deterministic way for a period of time, in which the total workload stays relatively constant. This deterministic behavior in the shifting of the workload manifests itself in the perfect correlations between the successive station times. As such, the results rigorously proven in the present subsection support the validity of the partially conjectured results in [66; 67]. □

Remark 3.2.9 Corollary 3.2.5 deals with the correlation of successive station times within a single cycle. This result can, however, be extended to correlations of successive times not belonging to the same cycle by modifying (3.16) in an obvious way. Hence, we obtain that the station time θ_i of queue i is perfectly correlated with the station time θ_{i+n} of queue $i + n$, i.e.,

$$\mathbb{C}\mathrm{or}[\theta_i^*, \theta_{i+n}^*] = 1, \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots. \tag{3.83}$$

□

3.2.4 Managerial insights

As described in Chapter 1, the SELSP is a common problem in process industries, where the utilization of capacity is typically extremely high. Thus, the heavy-traffic results of the present section can be used to get fundamental new insights into the behavior and performance of exhaustive lot-sizing policies in process industries. In particular, numerical tests in [P4; P5] show that the correlations among production runs for exhaustive lot-sizing policies are relatively high even in moderate traffic load. This not only implies that the cycle lengths are highly variable, but also that the system may drift away from average behavior for a significant period of time. Both effects may lead to higher inventory levels and costs at the production facility itself, are undesirable from an organizational point of view and hamper short-term decision making. Below, we elaborate further on these idiosyncrasies of exhaustive lot-sizing policies revealed by our heavy-traffic results.

First, the high correlations between production runs result in very long cycles from time to time, which increases the amount of safety stocks needed at the production facility and the concomitant holding costs. Bounding the production runs, and thus the cycle lengths, would be an effective strategy to solve this issue in environments with moderate or high capacity utilization (see, also, Chapter 4). However, we should bear in mind that in situations with extremely high load this bounding may be infeasible since it may lead to instability of some of the queues or even of the whole system.

Second, the high variance in cycle lengths caused by the correlations provides breeding ground for the conjecture that exhaustive lot-sizing policies do not lead to stability, regularity and discipline on the work floor. These properties are desirable from an organizational point of view, since they facilitate maintenance scheduling, workforce planning, purchasing of raw material, scheduling of subsequent processes and shipment of finished products (see, e.g., Chapter 1 of Van Nyen [172]). Schmidt *et al.* [191] report on a real-life case, where they actually observe the organizational flaws of exhaustive lot-sizing policies in a make-to-order production environment. By replacing the exhaustive policy in the plant by a strategy which stabilizes the cycle lengths, many direct and indirect improvements could be observed. Our heavy-traffic results can be seen as theoretical explanation of the lack of stability and discipline of exhaustive lot-sizing policies as observed in practice.

Third, the strong correlations between the lengths of production runs in heavy traffic prove that the performance of exhaustive lot-sizing policies in terms of, e.g., delivery times or WIP fluctuates strongly over time which may hinder short-term decision making. That is, the actual performance of the system is better than average for some periods of time, but for other periods the performance is below average. As Stoop [199] mentions, in the latter periods managers tend to make nervous myopic decisions in an attempt to reach average performance as quickly as possible, which may result in additional costs and lower long-term performance. The present subsection is closed with a remark.

Remark 3.2.10 Although the main motivation for our interest in correlations of station times is application oriented, these correlations are of theoretical interest as well. For example, they give an indication of the lengths of simulation runs needed to obtain sufficiently narrow confidence intervals of performance measures. The higher the correlations, the longer the simulation should be. As such, the results of the present section provide a theoretical explanation for the inefficiency of simulation techniques for (k-limited) polling systems as observed by, e.g., Blanc [41]. □

3.3 Increasing setup times

The present section, which is a strongly condensed version of [P10], presents an exact asymptotic analysis of the waiting time distribution in exhaustive polling systems when the *deterministic* setup times tend to infinity. To the best of our knowledge there exist only three papers in the vast polling literature addressing the problem of large setup times. [163] explores the descendant set approach in combination with the strong law of large numbers for renewal reward processes to analyze polling systems with deterministic setups and mixtures of exhaustive and gated service. [173] presents a somewhat simpler analysis in the case of an exhaustive system with deterministic setups, where the order of service is determined by a polling table. These results are generalized by the author in [P11] to, among other things, the complete class of policies allowing for a multi-type branching process interpretation (see Chapter 2). The main result in all of these papers is the fact that the scaled waiting time distribution converges in distribution to a uniform distribution.

The objective of the present section is the development of a new approach to derive the scaled waiting time distribution for polling systems with increasing deterministic setups. The main building block of our analysis is the MVA for polling systems, which shows that the scaled intervisit times converge in probability to a constant as the setup times increase to infinity. This result immediately leads to the known asymptotic expression for the scaled waiting time distribution. As such, an approach originally developed for the computation of *mean* waiting times (MVA) is straightforwardly applied to derive the asymptotic result for the complete waiting time *distribution*.

The remainder of the present section is organized as follows. Subsection 3.3.1 presents the main theorem, while subsequently Subsection 3.3.2 is mainly devoted to the proof of this theorem. Lastly, Subsection 3.3.3 argues that our results deepen the understanding of the behavior and performance of exhaustive lot-sizing policies in production environments with significant setup times.

3.3.1 Main result

The performance measure of interest is the waiting time W_i of a type-i customer, $i = 1, 2, \dots, N$, in case the *deterministic* setup times tend to infinity. Since the waiting time grows to infinity in the limiting case, we focus on the asymptotic scaled waiting time $W_i/\mathbb{E}[S]$ as $\mathbb{E}[S] \to \infty$, where the ratios of the setup times $a_i = \mathbb{E}[S_i]/\mathbb{E}[S]$ remain constant, $i = 1, 2, \dots, N$. By assuming that the mean of the scaled waiting time converges, it turns out that all limits in the present section are well defined.

The key result of the present section is the following theorem, where we take $\xrightarrow{p}$ to represent convergence in probability and $\xrightarrow{d}$ to represent convergence in distribution.

Theorem 3.3.1 *In case of deterministic setup times, we have for* $i = 1, 2, \dots, N$,

$$\frac{W_i}{\mathbb{E}[S]} \xrightarrow{d} W_i^*, \qquad (\mathbb{E}[S] \to \infty), \tag{3.84}$$

where W_i^* *is uniformly distributed on* $[0, \frac{1-\rho_i}{1-\rho}]$. □

An intuitive explanation of the uniform distribution emerging in the above theorem is that it represents the position of the server in the cycle, of which the length converges to a constant, on arrival of a tagged customer. With the help of Theorem 3.3.1 we can derive similar results for the PGF of the distribution of the scaled queue length $Q_i/\mathbb{E}[S]$ of queue i at arbitrary moments in time. That is,

$$\mathbb{E}[y^{\frac{Q_i}{\mathbb{E}[S]}}] = \mathbb{E}[e^{-\lambda_i \mathbb{E}[S](1-y^{\frac{1}{\mathbb{E}[S]}})\frac{W_i}{\mathbb{E}[S]}}] \xrightarrow{\mathbb{E}[S]\to\infty} \mathbb{E}[y^{\lambda_i W_i^*}], \tag{3.85}$$

where the first equality follows from application of the distributional form of Little's law [130] and the subsequent limit from the following standard limiting result,

$$\lim_{x\to\infty} x\left(1 - a^{\frac{1}{x}}\right) = -\ln(a). \tag{3.86}$$

We immediately observe from (3.85) that $Q_i/\mathbb{E}[S]$ equals $\lambda_i W_i^*$ in distribution as $\mathbb{E}[S] \to \infty$ implying that - although the individual service requests are discrete - the scaled queue length distribution converges to a *continuous* uniform distribution in the limit of increasing setup times as well. Intuitively, we can say that, when the setup times tend to infinity, the system behaves like a fluid model where customers keep trickling in and out like water. We come back to this issue in Remark 3.3.4.

From Theorem 3.3.1 we can perceive the following properties how the asymptotic scaled waiting time distribution depends on the system parameters.

Property 3.3.2 *For $i = 1, 2, \ldots, N$,*

1. *W_i^* is independent of the visit order;*

2. *W_i^* depends on the arrival rate and service time distribution of queue i only through the occupation rate ρ_i;*

3. *W_i^* depends on the other queues only through the total occupation rate ρ.*

□

Finally, throughout the present section, for each variable x its scaled counterpart $x/\mathbb{E}[S]$ as $\mathbb{E}[S] \to \infty$ is denoted by x^*.

3.3.2 Proof of main result

In order to prove Theorem 3.3.1, we use the following well-known relation between the waiting time W_i and the residual intervisit time R_{I_i} at queue i (see, e.g., [203]),

$$W_i = R_{I_i} + W_i^{M/G/1}, \qquad i = 1, 2, \ldots, N, \tag{3.87}$$

where $W_i^{M/G/1}$ is the waiting time in an $M/G/1$ queue with arrival rate λ_i and service time with first two moments $\mathbb{E}[B_i]$ and $\mathbb{E}[B_i^2]$. Note that R_{I_i} and $W_i^{M/G/1}$ are mutually independent. Since $W_i^{M/G/1}$ is *independent* of S, we can restrict ourselves to the intervisit times. To study these unknowns, we shift attention to the MVA set given by (3.13)-(3.16). As observed before, this set has no closed-form solution for general parameter settings. However, if the setup times are deterministic, we have $\mathbb{E}[S_i^2] = \mathbb{E}[S_i]^2$ and, thus, $\mathbb{E}[R_{S_i}] = \mathbb{E}[S_i]/2$, $i = 1, 2, \ldots, N$. If we now divide both sides of this set by $\mathbb{E}[S]$ and let $\mathbb{E}[S] \to \infty$, we obtain for $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, N-1$,

$$\sum_{n=1}^{N} q_{n,1}\mathbb{E}[Q_{i,n}^*] = \frac{\lambda_i}{1-\rho_i}\Big(\frac{(1-\rho)a_i^2}{2} + (1-q_{i,1})(\mathbb{E}[R_{\theta_{i+1,N-1}^*}] + a_i)\Big), \tag{3.88}$$

$$\sum_{n=i+1}^{i+j} \frac{q_{n,1}}{q_{i+1,j}}\mathbb{E}[Q_{i,n}^*] = \lambda_i\mathbb{E}[R_{\theta_{i+1,j}^*}], \tag{3.89}$$

$$\mathbb{E}[R_{\theta_{i,1}^*}] = \frac{1}{1-\rho_i}\left(\mathbb{E}[Q_{i,i}^*]\mathbb{E}[B_i] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\frac{a_i}{2}\right), \tag{3.90}$$

and for $j = 2, 3, \ldots, N-1$,

$$\mathbb{E}[R_{\theta_{i,j}^*}] = \frac{q_{i,1}}{q_{i,j}}\left(\frac{\mathbb{E}[R_{\theta_{i,1}^*}]}{\prod_{n=1}^{j-1}(1-\rho_{i+n})} + \sum_{n=1}^{j-1}\frac{a_{i+n} + \mathbb{E}[Q_{i+n,i}^*]\mathbb{E}[B_{i+n}]}{\prod_{m=n}^{j-1}(1-\rho_{i+m})}\right) + (1-\frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\theta_{i+1,j-1}^*}], \tag{3.91}$$

where we stress the similarities with the set (3.27)-(3.30) used in the proof of Lemma 3.1.3. Remark that terms like $q_{i,j}$ and $\mathbb{E}[S_i]/\mathbb{E}[\theta_{i,1}]$ represent fractions of time, obviously *independent* of (the limit of) S.

The scaled set (3.88)-(3.91) does have a closed-form solution given by (as can be verified by substitution),

$$\begin{aligned}
\mathbb{E}[Q^*_{i,i}] &= \frac{\lambda_i}{2}\left(\mathbb{E}[C^*] - \mathbb{E}[\theta^*_{i,1}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[C^*]\right), && i = 1,2,\ldots,N, && (3.92)\\
\mathbb{E}[Q^*_{i,i+n}] &= \lambda_i\left(\mathbb{E}[\theta^*_{i+1,n-1}] + \frac{1}{2}\mathbb{E}[\theta^*_{i+n,1}]\right), && i = 1,2,\ldots,N, && (3.93)\\
& && n = 1,2,\ldots,N-1, &&\\
\mathbb{E}[R_{\theta^*_{i,j}}] &= \frac{1}{2}\mathbb{E}[\theta^*_{i,j}], && i = 1,2,\ldots,N, && (3.94)\\
& && j = 1,2,\ldots,N-1, &&
\end{aligned}$$

where $\mathbb{E}[\theta^*_{i+1,0}] = 0$ and $\mathbb{E}[C^*] = \frac{1}{1-\rho}$ and $\mathbb{E}[\theta^*_{i,j}] = \sum_{n=i}^{i+j-1}(\frac{\rho_n}{1-\rho} + a_n)$. We refer to Remark 3.3.4 for an intuitive explanation of the solution (3.92)-(3.94).

It is easily verified from (3.94) that

$$\mathbb{V}\text{ar}[\theta^*_{i,j}] = 0, \qquad i = 1,2,\ldots,N, \quad j = 1,2,\ldots,N-1. \tag{3.95}$$

Since an immediate consequence of Chebyshev's inequality (see, e.g., [183]) is that a random variable with zero variance follows a deterministic distribution, Lemma 1 in [173] gives us, for $i = 1,2,\ldots,N$, and $j = 1,2,\ldots,N-1$,

$$\frac{\theta_{i,j}}{\mathbb{E}[S]} \xrightarrow{p} \mathbb{E}[\theta^*_{i,j}] = \sum_{n=i}^{i+j-1}(\frac{\rho_n}{1-\rho} + a_n), \qquad (\mathbb{E}[S] \to \infty). \tag{3.96}$$

By definition, this leads for $i = 1,2,\ldots,N$ to

$$\frac{I_i}{\mathbb{E}[S]} \xrightarrow{p} \sum_{n=i+1}^{i+N-1}(\frac{\rho_n}{1-\rho} + a_n) + a_i = \frac{1-\rho_i}{1-\rho}, \qquad (\mathbb{E}[S] \to \infty). \tag{3.97}$$

Subsequently, we recall the following well-known relation between the *cumulative distribution function* (CDF) $F_{I_i}(\cdot)$ of I_i and the *probability density function* (PDF) $f_{R_{I_i}}(\cdot)$ of R_{I_i} for $i = 1,2,\ldots,N$, (see, e.g., [183])

$$f_{R_{I_i}}(\cdot) = \frac{1 - F_{I_i}(\cdot)}{\mathbb{E}[I_i]}, \tag{3.98}$$

which, in combination with the fact that convergence in probability implies convergence in distribution, leads to for $i = 1,2,\ldots,N$,

$$\frac{R_{I_i}}{\mathbb{E}[S]} \xrightarrow{d} \frac{1-\rho_i}{1-\rho}U, \qquad (\mathbb{E}[S] \to \infty), \tag{3.99}$$

where U is a uniform random variable on $[0,1]$. Using (3.87) and recalling that $W_i^{M/G/1*} \xrightarrow{p} 0$ as $\mathbb{E}[S] \to \infty$ completes the proof. □

We close the present subsection with some remarks.

Remark 3.3.3 The astute reader has already noticed that we use the assumption of deterministic setup times only in the derivation of the closed-form solution (3.92)-(3.94) of the scaled MVA equations (3.88)-(3.91). In fact, we know that the MVA equations form a set of linear equations, where the $\mathbb{E}[R_{S_i}]$ only show up in the right-hand sides. Due to continuity of linear transformations, we can extend the result of Theorem 3.3.1 - and, thus, the results of [163] and [173] - by replacing the assumption of deterministic setup times by the following, less restrictive, one that $\mathbb{E}[R_{S_i}]$ approaches $\frac{1}{2}\mathbb{E}[S_i]$ as $\mathbb{E}[S] \to \infty$, $i = 1, 2, \dots, N$. Since $\mathbb{E}[R_{S_i}]$ can be rewritten as, for $i = 1, 2, \dots, N$,

$$\mathbb{E}[R_{S_i}] = \frac{1}{2}\mathbb{E}[S_i]\Big(1 + \frac{\mathbb{V}\mathrm{ar}[S_i]}{\mathbb{E}[S_i]^2}\Big), \tag{3.100}$$

this means that the *variance* of the setup times should remain *finite* in the limiting case of *infinite* setup times (or, at least, $\frac{\mathbb{V}\mathrm{ar}[S_i]}{\mathbb{E}[S_i]^2} \to 0$ as $\mathbb{E}[S] \to \infty$). □

Remark 3.3.4 The explicit solution of the MVA equations represented by (3.92)-(3.94) has an intuitively appealing interpretation, certainly worth mentioning. That is, in the case of increasing deterministic setup times the polling system converges to a deterministic cyclic system with continuous deterministic service rates $1/\mathbb{E}[B_i]$ and continuous demand rates λ_i at queue i, $i = 1, 2, \dots, N$. This means that in the limit the customers arrive to the system and are served at constant rates with no statistical fluctuation whatsoever and that the scaled queue lengths can be seen as continuous quantities, see (3.85). The station times and conditional lengths emerging in (3.92)-(3.94) are precisely equal to the corresponding quantities in such a deterministic cyclic system. This explanation also clearly indicates the difficulties arising in a system with increasing *stochastic* setup times, since it is certainly not obvious how such a polling system behaves in the limit. □

Remark 3.3.5 The present section demonstrates that MVA makes the asymptotic analysis of the waiting time distribution strikingly simple in case the setup times tend to infinity. The underlying reason for this is the fact that MVA explicitly gives, as by-product, the second moments of the station times, intervisit times and cycle lengths. In the case of deterministic setup times tending to infinity, MVA in particular reveals that the scaled intervisit times converge in probability to a constant immediately leading to the result of Theorem 3.3.1. □

Remark 3.3.6 The derivation of the results presented here, in particular the closed-form solution of the MVA equations, relies on the assumption that the setups are deterministic (see also Remark 3.3.3). However, in heavy traffic the impact of higher moments of the setup times on the waiting time distribution vanishes (see Section 3.2). By incorporating the heavy-traffic results for the variance of the intervisit times derived in Section 3.2, this implies that, as ρ tends to 1, the results of the present section can be easily extended to generally distributed setup times. That is, in case of general setup times, the distribution of the asymptotic scaled waiting time under heavy traffic reads, for $i = 1, 2, \dots, N$,

$$\frac{(1-\rho)W_i}{\mathbb{E}[S]} \xrightarrow{d} W_i^*, \qquad (\rho \uparrow 1 \quad \text{and then} \quad \mathbb{E}[S] \to \infty), \tag{3.101}$$

where W_i^* is uniformly distributed on $[0, 1-\rho_i]$ (see, also, Chapter 2). □

3.3.3 Managerial insights

We want to start this section by stressing the dissimilarities between the two sources of high capacity utilization - due to either high load or large setup times - in polling systems unearthed by the MVA framework. That is, in heavy traffic caused by high load a *diffusion* limit has turned out to apply, which implies that the (variable) *gamma* distribution is prevalent, for example, in the scaled cycle lengths and the marginal queue lengths at polling instants. In contrast, in the case of increasing setup times a *fluid* limit is obtained with a central role for the *deterministic* and *uniform* distributions revealing itself again, e.g., in the scaled cycle lengths and the marginal queue lengths. In particular, given the dissimilarities between the two asymptotic regimes revealed by the MVA framework, it is extremely important in practice to identify the actual source of high utilization of capacity in order to control and improve system performance.

We think that it is prudent to remind the reader that in Chapter 2 we have already presented results of the author on (branching-type) polling systems with large setup times. From a practical perspective, the most important result is - as shown in Theorem 2.2.9 - that the WIP level under the exhaustive lot-sizing policy is stochastically smaller than the WIP level under any alternative lot-sizing policy. More colloquially, this implies that the exhaustive lot-sizing policy is optimal, in terms of WIP levels, in systems with extremely large setup times and that, thus, production runs should not be bounded in such settings. This is, however, not true for general systems (with smaller setup times) where bounding of production runs can lead to significant improvement of system performance as we show in Chapter 4.

The author has analyzed practical cases of the SELSP in process industries, where the setup times were extremely large and deterministic. We believe that the results of the present section give, therefore, new and fundamental insights into the behavior and performance of (exhaustive) lot-sizing policies in process industries. In particular, we have shown that, in the case of increasing deterministic setup times, the polling system converges to a deterministic cyclic system. One would expect that, in practice, production managers rely more on rigid (deterministic) production strategies in case of large setups than they do in case of small setups in which dynamic (stochastic) policies are expected to be seen. This observation gives birth to two interesting lines of research.

First, the deterministic counterpart of the SELSP, the so-called ELSP - which has been proven to be NP-hard (see Hsu [121]) - has received lots of attention in the literature over the past decades (for surveys see, e.g., Elmaghraby [86] and Salomon [186]). Two major differences can be seen between production plans for the SELSP and the ELSP. First, a rigid cyclic production plan will not suffice anymore in a stochastic environment, since one has to be responsive to the dynamic changes in this environment. This means that dynamics have to be included in the production plan. Second, in a stochastic environment the inventories for the individual products play a more important role than in the deterministic case, as indicated by Sox *et al.* [195]. Inventories now do not only reduce the number of setups in a cycle, but they also serve as hedge against stock-outs and scheduling conflicts due to the variation in demand, production or setup times. It would be highly interesting to theoretically investigate whether in the limit of increasing setup times the difference between the SELSP and the ELSP for a large class of policies indeed vanishes and whether, consequently, theoretical results obtained for the latter - especially, for the sequencing decision - could be applied to the SELSP.

Motivation of this first research question is provided by numerical tests in Olsen [173], which show that limit theorems as obtained here for polling systems carry over to more general queueing systems. Moreover, Boxma *et al.* [50; 51] proposed a method to design effective polling tables - basically, solving the sequencing decision for fixed-sequence base-stock policies - based on a lower bound resulting from the deterministic counterpart of the polling systems. The accuracy of the method itself as well as the performance of the deterministic lower bound increases in systems with large setup times (see Olsen [173]). The results in the present section, obviously, give a strong theoretical foundation of these observations.

Second, we advocate a large-scale empirical practical study that investigates the main characteristics of production strategies in environments both with no - or negligible - setup times on the one hand and extremely large setup times on the other hand. In this way, one could survey the amount of dynamics incorporated into the production plan, which is hypothesized to have a negative correlation with the amount of setup times in a cycle. Although it is impossible to define a single proxy for the degree of dynamics, a possibility is to measure the ratio of the safety stocks and the total amount of stock in the system. As the total setup time in the system is increased from zero to infinity, based on the limiting results derived this ratio is conjectured to decrease from (almost) one to zero.

3.4 Other scheduling disciplines

Recall that in the SELSP the scheduling of orders within a queue is irrelevant due to the indistinguishability of the replenishment orders at the product facility and the inability to measure the production time of an order before the start of production (see Chapter 1). However, polling models have recently been used to study the Bluetooth and 802.11 protocols as well as scheduling policies at routers and i/o subsystems in web servers. In such settings, the system operator must decide on the order in which jobs within each queue are served. Since in many of these computer settings the workloads are known to have high variability, while it is often desirable to give different requests different priority levels in order to provide differentiated service, using a policy other than FCFS within the queues is appealing. Motivated by this application we now want to spend some time on the scheduling decision, which dictates how to schedule jobs within each queue. This section is based on part of the results in [P9].

Although both the number of papers analyzing polling systems and the number of papers analyzing scheduling policies are impressive, the combination of the two has received very little attention. There are only a few exceptions where the effect of priority-based policies is studied in polling systems, for example, [98; 192; 207; 210]. However, the results attained for such priority-based policies are mostly limited to pseudo-conservation laws and approximations that are exact only in special cases, e.g., symmetric polling systems. In the present section, we illustrate that the developed MVA framework allows the exact analysis of a variety of scheduling policies (many for the first time) implying that we can solve the scheduling decision to optimality.

The rest of the section is structured as follows. In Subsection 3.4.1 some additional notation is introduced. Subsection 3.4.2 deals with the MVA analysis of scheduling policies in polling systems with exhaustive service.

3.4.1 Notation

The number of jobs served during a visit to a queue is determined by the *exhaustive* service discipline. Then, during the visit to each queue, we allow jobs to be scheduled for service according to the m-class priority discipline or one of the disciplines summarized in Table 3.1. However, we limit the discussion to work-conserving disciplines. Our main interest is in the mean response time (sojourn time) $\mathbb{E}[T_i]$ of a type-i customer, $i = 1, 2, \dots, N$, which is defined as the time in steady state from a customer's arrival at queue i until the completion of his service. Often, it is more convenient to study the mean delay, $\mathbb{E}[D_i]$, which is defined as $\mathbb{E}[T_i] - \mathbb{E}[B_i]$. We stress that this mean delay $\mathbb{E}[D_i]$ does not necessarily have to equal the mean waiting time $\mathbb{E}[W_i]$. For clarity, we explicitly denote the dependency of these performance measures on the scheduling disciplines by superscripts.

3.4.2 Mean value analysis

To begin our study of scheduling in exhaustive polling systems, we consider the mean delay of a tagged arrival of size x, j_x, to queue i. When the tagged job arrives, it needs to wait at least until the server returns to queue i (a residual intervisit period), which equals, cf. (3.7),

$$\mathbb{E}[R_{I_i}] = \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]). \tag{3.102}$$

In addition to waiting $\mathbb{E}[R_{I_i}]$ before receiving service and to the job size x itself, de-

Scheduling disciplines	
FCFS	*First Come First Served* serves jobs in the order they arrive.
LCFS	*Last Come First Served* non-preemptively serves the job that arrived the most recently.
PLCFS	*Preemptive Last Come First Served* preemptively serves the most recent arrival.
SJF	*Shortest Job First* non-preemptively serves the job in the system with the smallest original size.
SRPT	*Shortest Remaining Processing Time* preemptively serves the job with the shortest remaining size.

Table 3.1: A brief description of the scheduling policies.

pending on the scheduling policy, the delay of j_x may include time devoted to serving (i) jobs that arrive after j_x begins service, (ii) jobs that arrived before j_x, (iii) jobs that arrived after j_x and before j_x receives service. We denote the contribution of the first piece as $c_1(x)$ and the second piece as $c_2(V_i)$, where V_i represents the stationary work at queue i. To simplify the computation of the third component, we notice that many common scheduling policies obey the following property:

Property 3.4.1 *The contributions to the delay of j_x from all jobs that arrive after j_x and before j_x receives service, denoted $c_3(B_i)$, are i.i.d. Further, once j_x receives service, no service is given to any other jobs that arrived before j_x.*

Many common policies obey Property 3.4.1, e.g., FCFS, LCFS, PLCFS, SJF and SRPT. Any policy which obeys Property 3.4.1 has the following representation for the mean delay $D_i(x)$ of a job of size x:

$$\begin{aligned}\mathbb{E}[D_i(x)] &= \mathbb{E}[c_1(x)] + \mathbb{E}[R_{I_i}] + \mathbb{E}\left[\sum_{j=1}^{N_A(R_{I_i})} \theta_{c_3(B_i^{(j)})}(c_3(B_i^{(j)}))\right] + \mathbb{E}[\theta_{c_3(B_i)}(c_2(V_i))] \\ &= \mathbb{E}[c_1(x)] + \mathbb{E}[R_{I_i}]\left(1 + \frac{\lambda_i \mathbb{E}[c_3(B_i)]}{1-\lambda_i\mathbb{E}[c_3(B_i)]}\right) + \frac{\mathbb{E}[c_2(V_i)]}{1-\lambda_i\mathbb{E}[c_3(B_i)]} \\ &= \mathbb{E}[c_1(x)] + \frac{\mathbb{E}[R_{I_i}] + \mathbb{E}[c_2(V_i)]}{1-\lambda_i\mathbb{E}[c_3(B_i)]}, \qquad (3.103)\end{aligned}$$

where $N_A(Y)$ is the number of arrivals during time Y, $B_i^{(j)}$ is the job size of the jth arrival, and $\theta_{B_i}(Y)$ is the length of a busy period started by Y work where service requirements of arrivals have i.i.d. sizes B_i.

Using (3.103), we can now easily obtain formulae for the mean delay of a handful of common scheduling policies under exhaustive polling models.

FCFS. The mean delay $\mathbb{E}[D_i]$ of FCFS in exhaustive polling systems equals the mean waiting time $\mathbb{E}[W_i]$ and has, consequently, been obtained in Section 3.1, but it serves as a useful example of applying (3.103). In the case of FCFS, only arrivals before the tagged job contribute to the delay of the tagged job. Thus, $\mathbb{E}[c_1(x)] = 0$, $\mathbb{E}[c_2(V_i)] = \mathbb{E}[V_i] = \rho_i\mathbb{E}[R_{B_i}] + E[Q_i]^{FCFS}\mathbb{E}[B_i]$ and $\mathbb{E}[c_3(B_i)] = 0$, which gives

$$\mathbb{E}[D_i(x)]^{FCFS} = \mathbb{E}[R_{I_i}] + \mathbb{E}[V_i] = \mathbb{E}[R_{I_i}] + \rho_i\mathbb{E}[R_{B_i}] + E[Q_i]^{FCFS}\mathbb{E}[B_i], \qquad (3.104)$$

and, thus, with the help of Little's Law,

$$\mathbb{E}[D_i(x)]^{FCFS} = \frac{\mathbb{E}[R_{I_i}] + \rho_i\mathbb{E}[R_{B_i}]}{1-\rho_i}. \qquad (3.105)$$

In order to view this in terms of the mean residual cycle length $\mathbb{E}[R_{C_i}] = \mathbb{E}[R_{\theta_{i+1,N}}]$, we use the well-known result that (see, e.g., Boxma [47]),

$$\mathbb{E}[D_i(x)]^{FCFS} = \mathbb{E}[R_{C_i}](1-\rho_i), \tag{3.106}$$

where a cycle obviously starts at a departure epoch of the server from queue i. Via (3.104) and (3.105), it follows that

$$\mathbb{E}[R_{C_i}] = \frac{\mathbb{E}[R_{I_i}] + \mathbb{E}[V_i]}{1-\rho_i} = \frac{\mathbb{E}[R_{I_i}] + \rho_i\mathbb{E}[R_{B_i}]}{(1-\rho_i)^2}. \tag{3.107}$$

This identity turns out to be useful for other policies as well since all work-conserving policies have the same mean residual cycle length $\mathbb{E}[R_{C_i}]$. In the remainder of this section, we derive for each individual scheduling discipline an arrival relation in terms of $\mathbb{E}[R_{C_i}]$. In this way, we isolate the effects of the setup times and the dependencies between station times into one quantity, which allows us to perform qualitatively simple comparisons of the mean delays across all the scheduling disciplines. These mean residual cycle lengths can be computed numerically via the developed MVA framework, since MVA provides - as a by-product - the mean residual station times, and thus the mean residual cycle lengths, which are independent of the scheduling discipline (as long as it is work-conserving).

LCFS. For LCFS, we have $\mathbb{E}[c_1(x)] = 0$, $\mathbb{E}[c_2(V_i)] = \rho_i\mathbb{E}[R_{B_i}]$, and $\mathbb{E}[c_3(B_i)] = \mathbb{E}[B_i]$ and, thus,

$$\mathbb{E}[D_i(x)]^{LCFS} \;=\; \frac{\mathbb{E}[R_{I_i}] + \rho_i\mathbb{E}[R_{B_i}]}{1-\rho_i} = \mathbb{E}[R_{C_i}](1-\rho_i) = E[D_i(x)]^{FCFS}. \tag{3.108}$$

In fact, LCFS is not alone in having $\mathbb{E}[D_i]$ the same as FCFS. As in the $M/G/1$ queue, all non-preemptive policies that do not use size information have the same mean delay in exhaustive polling systems.

PLCFS. Moving beyond non-preemptive policies, let us now consider PLCFS. Since all arrivals after the tagged job contribute to the delay, we have $\mathbb{E}[c_1(x)] = \rho_i x/(1-\rho_i)$. Further, $\mathbb{E}[c_2(V_i)] = 0$ and $\mathbb{E}[c_3(B_i)] = \mathbb{E}[B_i]$, which gives,

$$\mathbb{E}[D_i(x)]^{PLCFS} = \frac{\rho_i x + \mathbb{E}[R_{I_i}]}{1-\rho_i} = \mathbb{E}[R_{C_i}](1-\rho_i) + \frac{\rho_i}{1-\rho_i}(x - \mathbb{E}[R_{B_i}]). \tag{3.109}$$

Thus, we can see that $\mathbb{E}[D_i(x)]^{PLCFS} \leq \mathbb{E}[D_i(x)]^{FCFS} \Leftrightarrow x \leq \mathbb{E}[R_{B_i}]$, which is the same relation as in the $M/G/1$ setting.

Extending the framework. Since determination of $\mathbb{E}[c_2(V_i)]$ for priority-based policies can be problematic, we need to extend the framework. To handle such policies we view $\mathbb{E}[c_2(V_i)]$ as the work in a "transformed" FCFS queue, which allows us to mimic the FCFS derivation. In particular, it can be verified that the following property holds under SJF, SRPT and many other priority-based policies.

Property 3.4.2 *The contribution $c_2(V_i)$ can be viewed as the work in a "transformed" FCFS system where jobs arrive according to a Poisson process with rate λ_i having i.i.d. sizes $c_2'(B_i)$ and a different (maybe dependent) stream of jobs may arrive while the server is idle following a general (maybe non-Poisson) process. The resulting stationary amount of remaining work of the job receiving service is denoted $c_2''(R_{B_i})$.*[1]

[1] Note that this quantity does not assume that there is a job at the server and, thus, is a function of the load as well as the service distribution.

Examples of transformed systems are given below, but let us first examine the implications of Property 3.4.2. That is, denote the number of jobs in the queue of the "transformed" system as Q_i' and the delay in the transformed FCFS queue as $D_i^{FCFS'}$. Recall that the mean delay in a FCFS queue equals the mean work in the system plus $\mathbb{E}[R_{I_i}]$, thus $\mathbb{E}[D_i]^{FCFS'} = \mathbb{E}[R_{I_i}] + \mathbb{E}[c_2(V_i)]$. Given a policy obeys Property 3.4.2, we can write

$$\mathbb{E}[D_i]^{FCFS'} = \mathbb{E}[R_{I_i}] + \mathbb{E}[c_2''(R_{B_i})] + \mathbb{E}[Q_i']\mathbb{E}[c_2'(B_i)], \tag{3.110}$$

which gives, using Little's law,

$$\mathbb{E}[D_i]^{FCFS'} = \frac{\mathbb{E}[R_{I_i}] + \mathbb{E}[c_2''(R_{B_i})]}{1 - \lambda_i \mathbb{E}[c_2'(B_i)]}. \tag{3.111}$$

Combining the above with (3.103) and (3.107) gives

$$\begin{aligned}
\mathbb{E}[D_i(x)] &= \mathbb{E}[c_1(x)] + \frac{\mathbb{E}[R_{I_i}] + \mathbb{E}[c_2''(R_{B_i})]}{(1 - \lambda_i \mathbb{E}[c_2'(B_i)])(1 - \lambda_i \mathbb{E}[c_3(B_i)])} \\
&= \mathbb{E}[R_{C_i}] \left(\frac{(1-\rho_i)^2}{(1 - \lambda_i \mathbb{E}[c_2'(B_i)])(1 - \lambda_i \mathbb{E}[c_3(B_i)])} \right) \\
&\quad + \left(\mathbb{E}[c_1(x)] - \frac{\rho_i \mathbb{E}[R_{B_i}] - \mathbb{E}[c_2''(R_{B_i})]}{(1 - \lambda_i \mathbb{E}[c_2'(B_i)])(1 - \lambda_i \mathbb{E}[c_3(B_i)])} \right). \qquad (3.112)
\end{aligned}$$

The form of (3.112) is quite illustrative. The first term captures the growth as a function of the mean residual cycle length and the second term captures the tradeoff between giving priority to jobs that arrived earlier versus jobs that arrived later. In addition, (3.112) illustrates an important comparison between the $M/G/1$ model and exhaustive polling systems. Recalling that $\mathbb{E}[D_i]^{FCFS} = \mathbb{E}[R_{C_i}](1 - \rho_i)$, we have that

$$\begin{aligned}
\mathbb{E}[D_i(x)] &= \mathbb{E}[D_i(x)]^{FCFS} \left(\frac{(1-\rho_i)}{(1 - \lambda_i \mathbb{E}[c_2'(B_i)])(1 - \lambda_i \mathbb{E}[c_3(B_i)])} \right) \\
&\quad + \left(\mathbb{E}[c_1(x)] - \frac{\rho_i \mathbb{E}[R_{B_i}] - \mathbb{E}[c_2''(R_{B_i})]}{(1 - \lambda_i \mathbb{E}[c_2'(B_i)])(1 - \lambda_i \mathbb{E}[c_3(B_i)])} \right). \qquad (3.113)
\end{aligned}$$

The important point about the above is that the contribution functions $c_i(\cdot), c_i'(\cdot)$ and $c_i''(\cdot)$ are independent of the polling system. So, the only place the polling system impacts (3.113) is through $\mathbb{E}[D_i(x)]^{FCFS}$. Thus, the qualitative relationships between the mean delay of policies that satisfy Properties 3.4.1 and 3.4.2 are insensitive to the underlying structure of the polling system and only depend on the fact that queues are served exhaustively. Note that the quantitative differences between policies do depend on the structure of the polling systems though, since the relative weights of the two terms in (3.113) depend on the magnitude of $E[D_i(x)]^{FCFS}$.

SJF. To analyze SJF, which optimizes the mean delay among all non-preemptive policies, consider a transformed FCFS queue where jobs of size $\geq x$ are only allowed to arrive at the moment they begin to receive service in the standard SJF queue. Thus, jobs of size $< x$ still obey a Poisson process but jobs with size $\geq x$ do not. The mean delay for the tagged job is the same in both of these queues. Thus, for SJF, we have that $\mathbb{E}[c_1(x)] = 0$, $\mathbb{E}[c_2'(B_i)] = \mathbb{E}[B_i 1_{[B_i<x]}]$, $\mathbb{E}[c_2''(R_{B_i})] = \rho_i \mathbb{E}[R_{B_i}]$, and $\mathbb{E}[c_3(B_i)] = \mathbb{E}[B_i 1_{[B_i<x]}]$. Applying (3.112) gives

$$\mathbb{E}[D_i(x)]^{SJF} = \frac{\mathbb{E}[R_{I_i}] + \rho_i \mathbb{E}[R_{B_i}]}{(1 - \rho_i(x))^2} = \mathbb{E}[R_{C_i}] \left(\frac{1-\rho_i}{1-\rho_i(x)} \right)^2, \tag{3.114}$$

where $\rho_i(x) = \lambda_i \mathbb{E}[B_i 1_{[B_i<x]}]$. Thus, we can see that $\mathbb{E}[D_i(x)]^{SJF} \leq \mathbb{E}[D_i(x)]^{FCFS} \Leftrightarrow \rho_i(x) \leq 1 - \sqrt{1-\rho_i}$, which also holds in the $M/G/1$ setting.

To obtain the overall mean delay of SJF, we can simply integrate (3.114) as follows

$$\mathbb{E}[D_i]^{SJF} = \mathbb{E}[R_{C_i}] \int_0^\infty \left(\frac{1-\rho_i}{1-\rho_i(x)} \right)^2 f_i(x) dx. \tag{3.115}$$

Unfortunately though, no closed-form solution is available for this integral, but it can be proved that $\mathbb{E}[D_i]^{SJF} \leq \mathbb{E}[D_i]^{FCFS}$.

SRPT. As in the $M/G/1$ setting, SRPT optimizes the mean delay in exhaustive polling systems. However, the mean delay of SRPT has not been derived in this setting. In the case of SRPT, the transformed system that we use has jobs with original size $< x$ arrive at the same instants as normal, but has jobs with original size $\geq x$ arrive to the server at the moment they obtain remaining size x. Thus, they always arrive when the transformed system is idle. Thus, we obtain $\mathbb{E}[c_2'(B_i)] = \mathbb{E}[B_i 1_{[B_i<x]}]$ and $\mathbb{E}[c_2''(R_{B_i})] = \hat{\rho}_i(x)\mathbb{E}[R_{\min(B_i,x)}]$, where $\hat{\rho}_i(x) = \lambda_i \mathbb{E}[\min(B_i, x)]$. Further, noting that new arrivals contribute to the delay of the tagged job only when they are smaller than the remaining size of the tagged job, we have $\mathbb{E}[c_3(x)] = \mathbb{E}[B_i 1_{[B_i<x]}]$ and $\mathbb{E}[c_1(x)] = \int_0^x (\frac{1}{1-\rho_i(t)} - 1) dt = \int_0^x \frac{\rho_i(t)}{1-\rho_i(t)} dt$, where $\frac{dt}{1-\rho_i(t)}$ should be interpreted as the mean length of a busy period started by dt work including all new arrivals of size $< t$. Applying (3.112) gives

$$\begin{aligned} \mathbb{E}[D_i(x)]^{SRPT} &= \int_0^x \frac{\rho_i(t)}{1-\rho_i(t)} dt + \frac{\mathbb{E}[R_{I_i}] + \hat{\rho}_i(x)\mathbb{E}[R_{\min(B_i,x)}]}{(1-\rho_i(x))^2} \\ &= \mathbb{E}[R_{C_i}] \left(\frac{1-\rho_i}{1-\rho_i(x)} \right)^2 + \int_0^x \frac{\rho_i(t)}{1-\rho_i(t)} dt - \\ &\qquad \frac{\rho_i \mathbb{E}[R_{B_i}] - \hat{\rho}_i(x)\mathbb{E}[R_{\min(B_i,x)}]}{(1-\rho_i(x))^2}. \end{aligned} \tag{3.116}$$

As with SJF, we can obtain the overall mean delay of SRPT by integrating (3.116); however, such integration must be done numerically. But, without resorting to numerics, it is already evident that SRPT can provide significant reductions in mean delay when compared to FCFS and even SJF.

***m*-class priority queues.** We now move to the m-class priority discipline, where we limit our discussion to the non-preemptive variant. The mean delay of a class j job, $\mathbb{E}[D_i^{(j)}]$, is again easily derived from (3.112). Foregoing the details since they parallel the analysis of SJF, we have that $\mathbb{E}[c_1(B_i)] = 0$, $\mathbb{E}[c_2'(B_i)] = \mathbb{E}[B_i^{(k)} 1_{[k \leq j]}]$, $\mathbb{E}[c_2''(B_i)] = \rho_i \mathbb{E}[R_{B_i}]$, and $\mathbb{E}[c_3(B_i)] = \mathbb{E}[B_i^{(k)} 1_{[k<j]}]$. Thus, (3.112) gives

$$\mathbb{E}[D_i^{(j)}] = \mathbb{E}[R_{C_i}] \left(\frac{(1-\rho_i)^2}{(1-\sum_{k<j} \rho_i^{(k)})(1-\sum_{k\leq j} \rho_i^{(k)})} \right), \tag{3.117}$$

where $\rho_i^{(j)} = \lambda_i^{(j)} \mathbb{E}[B_i^{(j)}]$. Notice that the mean delay of SJF can be obtained by taking the appropriate limits. From (3.117) we can calculate the overall mean delay using

$$\mathbb{E}[D_i] = \sum_j \frac{\lambda_i^{(j)}}{\lambda_i} \mathbb{E}[D_i^{(j)}]. \tag{3.118}$$

Although this formula is easy to write, it hides the behavior of the mean delay as a function of the job sizes of each class. Since it is straightforward to show that the mean delay is minimized when priority is given to the classes that have small service requirements, it makes sense to consider threshold based policies. However, in general we cannot derive a closed-form expression for the optimal thresholds except in the case of two priority classes as shown below.

In this case, we can simplify the expression for the mean delay. In particular, letting t be the threshold used by the policy, we have

$$\frac{\mathbb{E}[D_i]}{\mathbb{E}[D_i]^{FCFS}} = \frac{\lambda_i^{(1)}}{\lambda_i}\frac{1-\rho_i}{1-\rho_i^{(1)}} + \frac{\lambda_i^{(2)}}{\lambda_i}\frac{1}{1-\rho_i^{(1)}} = \frac{1-\rho_i F_i(t)}{1-\rho_i^{(1)}}. \tag{3.119}$$

Differentiating this expression, we find

$$\frac{d}{dt}\left(\frac{\mathbb{E}[D_i]}{\mathbb{E}[D_i]^{FCFS}}\right) = \frac{-\rho_i f_i(t)(1-\rho_i^{(1)}) + \lambda_i t f_i(t)(1-\rho_i F_i(t))}{(1-\rho_i^{(1)})^2}, \tag{3.120}$$

which gives that the mean delay is minimized, when the threshold satisfies

$$\frac{t}{\mathbb{E}[B_i]} = \frac{1-\lambda_i \int_0^t s f_i(s) ds}{1-\rho_i F_i(t)}. \tag{3.121}$$

Although this expression is not explicit, it can be solved easily in the case of many common service distributions (see [P9]). Furthermore, notice that the optimal threshold is $\geq \mathbb{E}[B_i]$ for all service distributions (note that the optimal threshold is an increasing function of λ_i and as $\lambda_i \to 0$, $t \to \mathbb{E}[B_i]$) and depends on the shape of the distribution.

In [P9] we have extended the analysis of scheduling disciplines within the MVA framework to the case of gated service. One of the most striking observations provided by this framework is the fact that a large class of scheduling policies behaves the same in exhaustive polling models as in the standard $M/G/1$ model as seen in the present section, whereas scheduling policies in gated polling models have a very different effect than in the $M/G/1$ model. This difference manifests itself not only in the complexity of the analysis, but also in the impact a scheduling discipline has on the overall mean delay. Furthermore, [P9] presents some simple numerical experiments illustrating the performance of scheduling policies in exhaustive and gated polling systems and shows that the impact of scheduling within queues can be dramatic. One could postulate the (perhaps intuitively appealing) claim that scheduling within a queue has only a minor effect on overall system performance. Namely, one could argue that such a local decision only influences a small part of the delay of a customer, since a major part consists of the time until the server returns to the queue under consideration, which is unaffected by the scheduling policy. The results in [P9] refute this assertion. The explanation for this is that at an polling instant there is often a large batch of jobs waiting for service and, thus, the order in which these jobs are served really matters. We close this section with a remark.

Remark 3.4.3 Although we have seen that many common policies obey Properties 3.4.1 and/or 3.4.2, there are also policies that do not satisfy them. Foremost, all processor sharing (PS) type policies such as discriminatory, weighted, and multi-level PS do not satisfy either 3.4.1 or 3.4.2, implying that our analytic framework does not apply to these policies. In fact, these policies are fundamentally more difficult to analyze in exhaustive polling systems than they are in the $M/G/1$ model. To see this, notice that an analysis of the mean delay of PS in exhaustive polling systems depends on understanding the transient behavior of the queue length distribution under PS in the $M/G/1$ model, which is known to be a very difficult problem [131]. Thus, we leave the analysis of PS-type policies as an open question and note that, unlike policies that satisfy Properties 3.4.1 and/or 3.4.2, the behavior of PS is very different than it is in the stationary $M/G/1$ setting. However, not every policy that violates Properties 3.4.1 and/or 3.4.2 is difficult to analyze in exhaustive polling systems. In particular, the foreground-background policy violates these properties but can be analyzed directly. □

3.5 Conclusion and possible extensions

In the present chapter, we have introduced an MVA framework for the computation of the mean delays in exhaustive-type or gated-type polling systems. Without seriously complicating the analysis, MVA can be carried over to variants of the considered polling systems: (i) systems with Poisson batch arrivals, (ii) systems with fixed polling tables and (iii) discrete-time polling systems. Extensions to other polling systems, either in an exact or approximate way, are not inconceivable as well.

In particular, an interesting, and challenging, topic for further research would be the classification of a general class of polling systems for which the MVA framework can be applied. More specifically, one could examine whether the MVA approach could be extended to the complete class of branching-type policies. However, it turns out that there does not exist a one-to-one correspondence between the branching and the MVA class. That is, the author has successfully applied the MVA technique to the following models which do *not* satisfy the branching property: a two-queue 1-limited/exhaustive system, cf. [116], a cyclic polling system with the globally-gated service policy, cf. [52], a closed polling system with a fixed number of permanent customers, cf. [34] and a hybrid polling system with both permanent and transient customers, cf. [35]. For the sake of presentation, we omit these derivations, but they are available from the author upon request.

Given the sparsity and structure of the coefficient matrix of the MVA set it is not inconceivable that an efficient (iterative) algorithm can be developed to solve this set. Furthermore, using a software package for symbolic formula manipulation, the same structure and sparsity allow an explicit solution for moderate values of N, where we have to remark that in the open literature these solutions have only been published for very small values of N. Although the resulting solutions will, without doubt, be very cumbersome expressions, they may be used to gain additional theoretical insights into the behavior of the system. They could, for example, be used to give error bounds on and speed of convergence of the asymptotic heavy-traffic expressions for the mean delay as obtained in Section 3.2.

Chapter 4

Bounding production runs

The present chapter aims to get us a step nearer to answering the question raised in Chapter 1: What is the gain in performance of bounding production runs by means of the quantity-limited lot-sizing policy in multi-item production-inventory systems? We start this investigation with an *exact* analysis of a two-queue k-limited system, after that we move to the *approximate* analysis of a multi-queue k-limited system and, eventually, we conclude with *simulative* numerical evaluations on the k-limited policy. Since in the SELSP - and in various other applications of polling systems - the objective function oftentimes depends not only on the *mean* queue lengths but on the *complete* marginal queue length distributions, the main interest of the present chapter is in these distributions.

4.1 Systems with two queues

The present section, which is an abridged version of the manuscripts [P14; P18], considers a two-queue *state-dependent* polling model, in which a setup is incurred for a queue only when it is non-empty. In this model, the single server serves the high-priority queue *exhaustively* and the low-priority queue according to the *k-limited* service strategy. Throughout the present monograph, we have assumed that the machine incurs a setup for a certain product even when there is no shortfall for this product, which is of course suboptimal, as argued by Sox *et al.* [195]. This observation makes the practical relevance of the inclusion of state-dependent setups in the studied queueing model evident.

As comprehensively expounded in Chapter 2, for general k, an exact evaluation for the queue length distribution is only known in two-queue exhaustive/k-limited systems with *zero setup times* (see Lee [145] and Ozawa [176; 177]). In many applications such as the SELSP, however, the setup times may be substantial and the presence of these setup times may be crucial for the operation of the system. Nearly all of the existing literature on polling systems makes the assumption of state-independent setups. Notable exceptions are the recent studies of Altman *et al.* [32], Günalay and Gupta [117], Gupta and Srinivasan [118] and Singh and Srinivasan [193], where exhaustive-type and gated-type service disciplines are explored in combination with state-dependent setups. The choice of modeling state-independent setups is generally not motivated by an application but by the tractability of the resulting analysis. The present section is the first study combining the k-limited service discipline and (state-dependent) setups.

We think that it is appropriate that we also bring the paper of Borst *et al.* [43] to the attention, in which approximate optimal values of the service limits with respect to a weighted sum of mean waiting times are obtained for general k-limited polling systems with state-independent setup times. Of particular interest to the present section is the fact that they derive a (partially conjectured) rule stating that for optimal operation of these systems the queues with the highest priority (i.e., the queues with the highest value of the ratio between cost factor and mean service time) must have their service limit set at infinity. In a two-queue system this would result into the priority rule of the present section providing additional evidence for the significance of the present study.

The main contribution of the present section is two-fold. First, the model in [145] is generalized by including state-dependent setups. In particular, we obtain the transforms of the queue length and sojourn time distributions under the assumption of Poisson arrivals, generally distributed service times and generally distributed setup times. Second, we demonstrate how the results of the analysis can be applied in the evaluation of a two-item instance of the SELSP. We observe significant cost reductions by application of the k-limited policy, compared to the standard exhaustive policies, in such settings indicating the potential of the k-limited service discipline as priority rule in production environments.

The rest of the present section is organized as follows. In Subsection 4.1.1, we present the model description including the stability conditions and the balance equations. Subsection 4.1.2 derives the PGFs of the joint queue length distributions both at service completion epochs and at arbitrary instants. The penultimate subsection is devoted to the application in the SELSP. Some concluding remarks are presented in Subsection 4.1.4.

4.1.1 Model description

We start this subsection with the specific notation and assumptions used throughout the present section. Then, we present the state description together with the corresponding balance equations.

Notation and assumptions. Consider a two-queue case of the basic polling system as described in Chapter 2, in which the high-priority queue 1 is served *exhaustively*, whereas the low-priority queue 2 is served according to the *k-limited* strategy. The combination of these service policies obviously creates a preferential treatment of type-1 customers. Unlike the basic polling system of Chapter 2, the setup times are assumed to be *state-dependent*, i.e., the server incurs a setup for a queue only when it is non-empty. When both queues are empty, the server stops working. He starts again upon arrival of a new customer and, then, he has to setup irrespective of the type of the last customer served before the idle time. Moreover, if the server has served k customers of the low-priority queue and the high-priority queue is empty, the server starts a new sequence up to k customers of this low-priority class after a new setup time of the low-priority queue. However, in this case no setup is incurred for the high-priority queue (which is the standard assumption in polling systems with state-independent setup times). It is important to note that the analysis of the present section also fully holds in the case of state-independent setups, *mutatis mutandis.*

We define $U_i(z_1, z_2)$ and $R_i(z_1, z_2)$ as the PGFs of the number of type-1 and type-2 arrivals during a service time and a setup time at queue i, respectively. That is,

$$U_i(z_1, z_2) = \beta_i(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)), \qquad i = 1, 2, \tag{4.1}$$

$$R_i(z_1, z_2) = \sigma_i(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)), \qquad i = 1, 2. \tag{4.2}$$

The quantities

$$r_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \quad \text{and} \quad r_2 = \frac{\lambda_2}{\lambda_1 + \lambda_2}, \tag{4.3}$$

denote the probabilities that the server switches to queue 1 and 2 after an idle period, respectively.

Winands *et al.* [P18] heuristically derive the stability conditions for the two queues by deploying arguments similar to those used by Ibe and Cheng [123], who study the stability conditions for a variety of polling systems without state-dependent setups. Furthermore, in [P18] the heuristic derivation has been proved rigourously by using fluid limit methodology described by Dai [75] and Dai and Meyn [76] and applied to polling models by Down [81] (see the theorem below).

Theorem 4.1.1 *The two-queue system is stable if and only if*

$$\hat{\rho} = \rho_1 + \rho_2 + \frac{\lambda_2}{k}\left(\mathbb{E}[S_2] + r\mathbb{E}[S_1]\right) < 1, \tag{4.4}$$

where r represents the probability that the number of type-1 arrivals during a setup time plus k successive service times at queue 2 is not equal to zero, i.e.,

$$r = 1 - R_2(0,1)(U_2(0,1))^k. \tag{4.5}$$

Proof. See Winands *et al.* [P18]. □

State description and balance equations. We study the system at embedded epochs of service completions of customers. The state of the system $\mathbf{Q}(n)$ just after the n^{th} departure from the system can be described by the following three variables:

1. $Q_1(n)$: the number of customers in queue 1;
2. $Q_2(n)$: the number of customers in queue 2;
3. $C(n)$: equals zero when the n^{th} departure is a type-1 customer, while it equals the number of type-2 departures since the last setup when the n^{th} departure is a type-2 customer.

The associated stochastic process,

$$\mathbf{Q}(n) = \{\left(Q_1(n), Q_2(n), C(n)\right), n = 1, 2, \ldots\}, \tag{4.6}$$

is an aperiodic and irreducible three-dimensional Markov chain. Let

$$\pi(q_1, q_2, c) = \lim_{n\to\infty} P[\left(Q_1(n), Q_2(n), C(n)\right) = (q_1, q_2, c)], \tag{4.7}$$

be the limiting probability and define the corresponding generating functions for this Markov chain as follows

$$p_1(z_1, z_2) = \sum_{q_1=0}^{\infty}\sum_{q_2=0}^{\infty} \pi(q_1, q_2, 0) z_1^{q_1} z_2^{q_2}, \tag{4.8}$$

$$p_{2,j}(z_1, z_2) = \sum_{q_1=0}^{\infty}\sum_{q_2=0}^{\infty} \pi(q_1, q_2, j) z_1^{q_1} z_2^{q_2}, \qquad j = 1, 2, \ldots, k. \tag{4.9}$$

Now, the following set of $k+1$ balance equations for equally many unknowns $p_1(z_1, z_2)$ and $p_{2,j}(z_1, z_2)$, $j = 1, 2, \ldots, k$, holds

$$\begin{aligned} p_1(z_1, z_2) &= \frac{U_1(z_1, z_2)}{z_1}\Big\{p_1(z_1, z_2) - p_1(0, z_2) + \\ &\Big[c_0 r_1 z_1 + \sum_{j=1}^{k-1}[p_{2,j}(z_1, 0) - p_{2,j}(0, 0)] + p_{2,k}(z_1, z_2) - p_{2,k}(0, z_2)\Big] R_1(z_1, z_2)\Big\}, \end{aligned} \tag{4.10}$$

$$p_{2,1}(z_1, z_2) = \frac{U_2(z_1, z_2) R_2(z_1, z_2)}{z_2}\alpha(z_2), \tag{4.11}$$

$$p_{2,j}(z_1, z_2) = \frac{U_2(z_1, z_2)}{z_2}\Big\{p_{2,j-1}(z_1, z_2) - p_{2,j-1}(z_1, 0)\Big\}, \qquad j = 1, 2, \ldots, k, \tag{4.12}$$

with

$$c_0 = p_1(0,0) + \sum_{j=1}^{k} p_{2,j}(0,0), \tag{4.13}$$

the probability that the system is left idle at a departure epoch and

$$\alpha(z_2) = c_0 r_2 z_2 + p_1(0,z_2) - p_1(0,0) + p_{2,k}(0,z_2) - p_{2,k}(0,0). \tag{4.14}$$

These balance equations are formulated by considering all the possible states at the previous departure epoch from which we can reach the current state. We explain (4.10), which describes the case that the current departure is a type-1 customer. First of all, the previous departure could be a type-1 departure which did not leave the system idle. This event corresponds to the term $p_1(z_1,z_2) - p_1(0,z_2)$. Secondly, the term $c_0 r_1 z_1$ represents the event that the previous departure left the system idle and that the first new arriving customer is of type 1. Thirdly, the term $\sum_{j=1}^{k-1}[p_{2,j}(z_1,0) - p_{2,j}(0,0)]$ represents the event that the last departure was a type-2 customer that was not the k^{th} in the sequence and that left queue 2, but not the complete system, idle. Finally, $p_{2,k}(z_1,z_2) - p_{2,k}(0,z_2)$ corresponds to the event that the last departure was a type-2 customer that was the k^{th} in the sequence and that queue 1 was not empty. The explanations of (4.11) and (4.12) are similar.

The function $\alpha(\cdot)$ is recognized as the generating function of the number of type-2 customers at moments a setup for queue 2 is initiated. Since at these specific points in time Q_1 and C are both equal to zero, the state description can be reduced to a single dimension represented by the function $\alpha(\cdot)$. The analysis of the next subsection is oriented towards relating the unknown generating functions $p_1(\cdot)$ and $p_{2,j}(\cdot,\cdot)$ to this function $\alpha(\cdot)$.

4.1.2 Exact analysis of queue lengths

This subsection presents the derivation of the generating functions of the joint queue length distributions at service completion epochs, which are, thereupon, used to derive expressions for the PGFs of the marginal queue size distributions at arbitrary instants.

Joint queue lengths at service completion epochs. To derive the generating functions of the joint queue length distributions at service completion epochs, we successively substitute (4.12) into itself and, then, into (4.11) which yields, for $j = 1, 2, \dots, k$,

$$p_{2,j}(z_1,z_2) = \frac{U_2^j(z_1,z_2)R_2(z_1,z_2)\alpha(z_2) - \sum_{l=1}^{j-1} z_2^{j-l} U_2^l(z_1,z_2) p_{2,j-l}(z_1,0)}{z_2^j}. \tag{4.15}$$

Notice that (4.15) gives an expression of $p_{2,j}(\cdot,\cdot)$, $j = 1, 2, \dots, k$, as a function of the unknown functions $\alpha(\cdot)$ and $p_{2,l}(\cdot,0)$, $l = 1, 2, \dots, j-1$.

Now, we turn our attention to $p_1(\cdot,\cdot)$. Substituting (4.15) for $j = k$ into (4.10) and using (4.13) and (4.14) gives us, after some straightforward manipulations,

$$\begin{aligned}\Big(z_1 - U_1(z_1,z_2)\Big)p_1(z_1,z_2) \;=\; & U_1(z_1,z_2)\Big\{c_0 r_1 z_1 R_1(z_1,z_2) + (R_1(z_1,z_2)-1)p_1(0,z_2) + \\ & \Big(\Big(\frac{U_2(z_1,z_2)}{z_2}\Big)^k R_2(z_1,z_2)\alpha(z_2) - \alpha(z_2) + c_0 r_2 z_2 + \\ & \sum_{j=1}^{k-1}\Big[1 - \Big(\frac{U_2(z_1,z_2)}{z_2}\Big)^j\Big] p_{2,k-j}(z_1,0) - c_0\Big) R_1(z_1,z_2)\Big\}.\end{aligned} \tag{4.16}$$

We eliminate $p_1(0,z_2)$ from the above equation by rewriting (4.14) as follows

$$\begin{aligned} p_1(0,z_2) &= \alpha(z_2) - c_0 r_2 z_2 + p_1(0,0) - p_{2,k}(0,z_2) + p_{2,k}(0,0) \\ &= \alpha(z_2) + c_0(1 - r_2 z_2) - \Big(\frac{U_2(0,z_2)}{z_2}\Big)^k R_2(0,z_2)\alpha(z_2) \qquad (4.17) \\ &\quad - \sum_{j=1}^{k-1}\Big[1 - \Big(\frac{U_2(0,z_2)}{z_2}\Big)^j\Big] p_{2,k-j}(0,0), \end{aligned}$$

which yields

$$\begin{aligned} \Big(z_1 - U_1(z_1,z_2)\Big)p_1(z_1,z_2) &= U_1(z_1,z_2)\Big\{(\frac{\beta(z_1,z_2)}{z_2^k} - 1)\alpha(z_2) + \\ &\quad R_1(z_1,z_2)\sum_{j=1}^{k-1}\frac{\delta_j(z_1,z_2)}{z_2^k}p_{2,k-j}(z_1,0) + \qquad (4.18) \\ &\quad D(z_1,z_2) - (R_1(z_1,z_2) - 1)\sum_{j=1}^{k-1}\frac{\delta_j(0,z_2)}{z_2^k}p_{2,k-j}(0,0)\Big\}, \end{aligned}$$

where

$$\begin{aligned} D(z_1,z_2) &= c_0\Big[r_1 z_1 R_1(z_1,z_2) + r_2 z_2 - 1\Big], \qquad (4.19) \\ \beta(z_1,z_2) &= U_2^k(z_1,z_2)R_1(z_1,z_2)R_2(z_1,z_2) - U_2^k(0,z_2)(R_1(z_1,z_2) - 1)R_2(0,z_2), \qquad (4.20) \\ \delta_j(z_1,z_2) &= z_2^k - z_2^{k-j}U_2^j(z_1,z_2). \qquad (4.21) \end{aligned}$$

It is again important to notice that via (4.18), $p_1(\cdot,\cdot)$ is also expressed as a function of the unknown functions $\alpha(\cdot)$ and $p_{2,j}(\cdot,0)$, $j = 1,2,\dots,k-1$.

It is well known that for each (fixed) $|z_2| \le 1$ the term $z_1 - U_1(z_1,z_2)$ in (4.18) has exactly one zero $z_1 = \xi(z_2)$ with $|z_1| \le 1$ if $\rho_1 < 1$. More specifically,

$$z_1 = \xi(z_2) = \theta_1[\lambda_2(1 - z_2)], \qquad (4.22)$$

where $\theta_1(\cdot)$ is the LST of the busy period of a standard M/G/1 queue with arrival rate λ_1 and LST of the service time $\beta_1(\cdot)$ (see, e.g., Takács [201]). Thus, $\xi(\cdot)$ can be seen as the PGF of the distribution of the number of type-2 arrivals during such an M/G/1 busy period.

Remark 4.1.2 It is interesting to note that the function $\beta(\xi(z),z)$ is the PGF of the number of type-2 customers arriving in a cycle for queue 2 in which the maximum of k customers is served. □

By analyticity of $p_1(z_1,z_2)$, the right-hand side of (4.18) should vanish when $z_1 = \xi(z_2)$. Hence,

$$\begin{aligned} \alpha(z) &= \frac{D(\xi(z),z)z^k + R_1(\xi(z),z)\sum_{j=1}^{k-1}\delta_j(\xi(z),z)p_{2,k-j}(\xi(z),0)}{z^k - \beta(\xi(z),z)} - \\ &\qquad \frac{(R_1(\xi(z),z) - 1)\sum_{j=1}^{k-1}\delta_j(0,z)p_{2,k-j}(0,0)}{z^k - \beta(\xi(z),z)}, \qquad (4.23) \end{aligned}$$

and $\alpha(z)$ is formulated as a function of the unknown functions $p_{2,j}(\cdot,0)$, $j=1,2,\ldots,k-1$.

To eliminate these unknown functions, we differentiate the numerator and denominator of (4.15) j times with respect to z_2 and, by L'Hospital's rule, we obtain the following recursion, for $j=1,2,\ldots,k-1$,

$$p_{2,j}(x,0)=\sum_{l=1}^{j}\frac{c_l\frac{d^{j-l}}{dy^{j-l}}[U_2^j(x,y)R_2(x,y)]\Big|_{y=0}}{(j-l)!}-\sum_{l=1}^{j-1}\frac{\frac{d^l}{dy^l}[U_2^l(x,y)]\Big|_{y=0}}{l!}p_{2,j-l}(x,0),\quad (4.24)$$

where c_l, $l=1,2,\ldots,k-1$, represent the probabilities that l type-2 customers are present at the start of a setup for this queue, i.e.,

$$c_l=\frac{\frac{d^l}{dy^l}[\alpha(y)]\Big|_{y=0}}{l!},\qquad l=1,2,\ldots,k-1.\qquad (4.25)$$

From this interpretation of c_l one can easily deduce a probabilistic interpretation of (4.24) as well, i.e., left and right hand side clearly represent the PGF of the number of customers in queue 1 when the server leaves queue 2 due to the fact that the latter queue is empty after j customers are served, $j=1,2,\ldots,k-1$. Of course, we could have derived (4.24) directly by using this probabilistic interpretation and, hence, avoid use of L'Hospital's rule.

By (4.24) we can write $p_{2,j}(\cdot,0)$ as a function of the unknown probabilities c_j, $j=0,1,\ldots,k-1$. Moreover, with the help of (4.15), (4.18) and (4.23) the generating functions $p_{2,j}(\cdot,\cdot)$, $p_1(\cdot)$ and $\alpha(\cdot)$ can be expressed in terms of these constants as well. The problem of finding these generating functions is, thus, reduced to finding the unknown probabilities c_j, which can be computed as follows. Given that (4.4) holds, the following theorem states that the denominator of (4.23) has exactly k zeros on or within the unit circle.

Theorem 4.1.3 *Under (4.4), it holds that $z^k=\beta(\xi(z),z)$ has k roots on or within the unit circle.*

Proof. The derivative of $\beta(\xi(z),z)$ at $z=1$ equals

$$\beta'(\xi(1),1)=\frac{\lambda_2}{1-\rho_1}(k\beta_2+\tau_1+\tau_2)-\frac{\lambda_2}{1-\rho_1}U_2^k(0,z_2)R_2(0,z_2)\tau_1=\frac{k\rho_2+\lambda_2(r\tau_1+\tau_2)}{1-\rho_1},\qquad (4.26)$$

where r is defined by (4.5). Because we have assumed (4.4), we obtain $\beta'(\xi(1),1)<k$ and the result follows from Theorem 4.A.4 in Appendix 4.A. □

Since $\alpha(z)$ is bounded in $|z|\leq 1$, the zeros in the numerator must be canceled by corresponding zeros in the denominator. One of the zeros equals one and leads to a trivial equation. However, the normalization condition provides an additional equation and, therefore, we have a set of k linear equations. By making the assumption that the k roots of $z^k=\beta(\xi(z),z)$ on or within the unit circle are all distinct (see Remark 4.1.4), this set of equations has a unique solution for c_j, $j=0,1,\ldots,k-1$. This completes the determination of the generating functions of the queue length distributions at service completion epochs and, hence, in the remainder of the present section we assume that these generating functions are known.

Remark 4.1.4 If one or more roots of $z^k=\beta(\xi(z),z)$ on or within the unit circle coincide, our reasoning needs to be slightly modified. That is, for $\alpha(z)$ to be bounded in $|z|\leq 1$, the numerator of (4.23) should still have the same zeros as the denominator of (4.23) *and* with the same multiplicity. Additional equations can, therefore, be obtained by requiring that the derivative(s) of the numerator should also vanish where the denominator has a zero of higher multiplicity. □

Marginal queue lengths at arbitrary instants. From the results of the previous subsection, we can obtain expressions for the PGF $q_i(\cdot)$ of the marginal queue size distributions of queue i at type-i departure epochs, i.e.,

$$q_1(z) = \frac{p_1(z,1)}{r_1}, \quad \text{and} \quad q_2(z) = \frac{\sum_{j=1}^{k} p_{2,j}(1,z)}{r_2}. \tag{4.27}$$

By using a standard level crossing argument, in combination with PASTA, it can be shown that the marginal queue length distribution of queue i at type-i departure epochs and at arbitrary instants in time denoted by L_i are the same. Hence, the PGFs for these marginal distributions are given by (4.27). From (4.27) we can easily obtain the LST $T_i(\cdot)$ of the sojourn time distribution of a type-i customer. Since the number of type-i customers left behind by a tagged type-i customer equals the number of customers arrived during the sojourn time of this tagged customer, we have

$$T_i(z) = q_i(1 - \frac{z}{\lambda_i}), \qquad i = 1, 2, \tag{4.28}$$

which is known as the distributional form of Little's law (see, e.g., Keilson and Servi [130]).

4.1.3 Application

In the present subsection, we use the analysis of the preceding subsection to numerically evaluate a stochastic two-item instance of the SELSP. Consider a two-product instance of the base-stock system described in Chapter 1, where for product 1 an exhaustive lot-sizing policy is implemented, while for product 2 a quantity-limited lot-sizing policy is used. Imitating the reasoning expounded in the first introductory chapter, we can argue that the shortfall distribution of product i at the production facility is identical to the queue length distribution of queue i in the queueing model of the present section. Hence, by the procedure presented in Subsection 4.1.2, the steady-state net stock level distribution for both products can be computed. Thereupon, assuming negligible setup costs and a linear cost structure for the holding cost h_i and penalty costs p_i for product i yields the optimal base-stock levels b_i^* (see, again, Chapter 1),

$$b_i^* = \min\{n \in \mathbb{N}_0 | P[L_i \leq n] \geq \frac{p_i}{p_i + h_i}\}, \qquad i = 1, 2, \tag{4.29}$$

and the concomitant costs $Z(b^*)$,

$$Z(b^*) = \mathbb{E}[\sum_{i=1}^{N} c_i\,(b_i^* - L_i)]. \tag{4.30}$$

Of course, a whole plethora of cases can be studied: different values of the quantity limit, choice of service time distributions and their parameters, choice of setup distributions and their parameters, different (ratios between) cost factors, etcetera. However, the aim of the present subsection is to present some illustrative cases which show the potential of the k-limited service policy in the context of the SELSP.

As described before, the PGFs derived in the previous subsection have to be finished off with a number of zeros, which are numerically computed by using the Chaudhry QROOT software package [64]. We, then, use the method presented by Abate and Whitt [24] to numerically invert these PGFs. Unfortunately, for large quantity limits, numerical problems have been encountered in the procedure. Therefore, we confine ourselves to cases with small quantity limits. In the numerical evaluation we also present results for the special case of $k = \infty$, which amounts to a two-product model with an exhaustive lot-sizing

Product information		
Parameter	**Product** 1	**Product** 2
Demand (rate)	Poisson(0.375)	Poisson(0.375)
Processing times	Exp(1.0)	Exp(1.0)
Setup times	Exp(0.25)	Exp(0.25)
Holding costs	10	1
Backlogging costs	90	9

Table 4.1: Product information for Case 1.

policy for both products. In this case, we do not use the procedure of Subsection 4.1.2, but we have implemented a discrete event simulation. Each simulation run is sufficiently long such that the widths of the 95% confidence intervals of the performance measures of interest are smaller than 1% of the estimated value.

The examination starts with an initial case, which is subsequently being perturbed into 6 cases to study the effect of (1) (ratio between) the loads, (2) (ratio between) cost factors, (3) setup times.

Case 1 (*Initial case*). Suppose that product 1 is a product with high costs, whereas product 2 is of secondary importance compared to the first product. Table 4.1 shows the detailed specifications for these two products (where the numbers between brackets are the means of the corresponding distributions). Table 4.2 shows the costs Z_i per product, the total costs Z and the optimal base-stock levels b_i^* as a function of the quantity limit k. Firstly, we observe that the costs for product 2 are decreasing in the quantity limit, whereas the costs for product 1 increase with this limit. Secondly, the optimal value of the quantity limit with respect to total costs is equal to 2. For smaller limits the amount of capacity available for production is too low, while larger limits lead to more variable cycle lengths. Thirdly, the optimal base-stock levels for product 2 are non-increasing in the quantity limit, while the optimal base-stock levels for product 1 are non-decreasing in this limit. Finally, it is interesting to compare this table with the total costs equalling 59.2 that would be incurred if a standard exhaustive policy were implemented for both products. Via a k-limited policy for product 2, we may, thus, save 35.0% compared to the latter policy clearly showing the advantage of the k-limited policy in a production environment.

Case 2 and 3 (*Effect of the load*). These cases are similar to the initial case, except that the demand rates are perturbed, i.e., in Case 2 the demand rates of product 1 and 2 equal 0.15 and 0.6, whereas in Case 3 product 1 and 2 have demand rates 0.6 and 0.15, respectively. Comparing the results in Tables 4.3 and 4.4 with the costs for exhaustive base-

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	3	27.5	13	13.9	41.4
2	*3*	*28.8*	*9*	*9.8*	*38.5*
3	3	30.5	7	8.7	39.2
4	3	32.5	7	8.1	40.6
5	3	34.6	6	7.7	42.3
6	4	36.5	6	7.4	43.9

Table 4.2: Case 1.

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	1	13.8	24	24.9	38.7
2	1	15.4	12	12.5	27.8
3	*1*	*17.1*	*10*	*10.6*	*27.7*
4	2	18.6	9	9.8	28.4
5	2	19.2	8	9.3	28.5
6	2	19.9	8	9.0	28.8

Table 4.3: Case 2.

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	*5*	*47.6*	*5*	*6.6*	*54.2*
2	5	48.7	4	5.8	54.5
3	5	49.9	4	5.4	55.3
4	5	51.0	4	5.2	56.2
5	5	52.0	4	5.0	57.0
6	5	53.0	3	4.9	57.9

Table 4.4: Case 3.

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	3	13.7	13	13.9	27.6
2	3	14.4	9	9.8	24.2
3	*3*	*15.2*	*7*	*8.7*	*24.0*
4	3	16.2	7	8.1	24.3
5	3	17.3	6	7.7	25.0
6	4	18.3	6	7.4	25.6

Table 4.5: Case 4.

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	3	5.5	13	13.9	19.4
2	3	5.8	9	9.8	15.5
3	3	6.1	7	8.7	14.8
4	*3*	*6.5*	*7*	*8.1*	*14.6*
5	*3*	*6.9*	*6*	*7.7*	*14.6*
6	4	7.3	6	7.4	14.7

Table 4.6: Case 5.

stock policies equalling 49.3 and 61.7, respectively, once more significant cost reductions are observed by application of the k-limited policy. Of course, the advantages of the k-limited policy are much more pronounced in case the low-priority product 2 has the highest demand rate.

Case 4 and 5 (*Effect of the cost factors*). In the fourth and fifth case, we consider systems similar to that of the initial case and perturb the cost factors. That is, the cost factors for product 2 remain unaltered, while the holding and penalty costs for product 1 are decreased to 5 and 45 for Case 4 and to 2 and 18 for Case 5, respectively. Although the same conclusions as in Case 1 can be drawn from Tables 4.5 and 4.6, it should be observed that the advantages of the k-limited discipline are a bit less pronounced in these cases due to the leveling of the costs among the products. In fact, application of exhaustive policies for both products, which is normally done, would lead to total costs of 32.3 and 16.1, respectively.

Case 6 and 7 (*Effect of the setup times*). In Cases 6 and 7 we examine what the effects are of the sizes of the setup times. We therefore study two cases similar to the initial case, but in which the mean setup times equal 0.5 and 0.1, respectively. See Tables 4.7 and 4.8 for the results. Notice that for Case 6 the system is not stable when the quantity limit is chosen to be equal to 1. In case we implement the exhaustive base-stock policy for both products, the total costs are given by 62.7 and 56.3. This leads to the intuitively appealing conclusion that the advantages of the k-limited service discipline slightly increase in case the setup times vanish, but we stress that even in the case of large setup times the k-limited strategy still shows its superiority over the exhaustive policy.

In the present subsection we have presented some cases, which lead to the conjectures that the widely used exhaustive policy is not the most effective strategy in (frequently encountered) asymmetric production situations as well as that it may be desirable that

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	–	–	–	–	–
2	3	30.2	16	16.9	47.1
3	*3*	*32.4*	*11*	*12.1*	*44.5*
4	4	34.7	9	10.4	45.0
5	4	36.0	8	9.4	45.4
6	4	37.6	7	8.8	46.4

Table 4.7: Case 6.

Output					
k	b_1^*	Z_1	b_2^*	Z_2	Z
1	*2*	*26.7*	*8*	*8.8*	*35.4*
2	3	28.3	7	7.9	36.2
3	3	29.8	7	7.5	37.2
4	3	31.5	7	7.2	38.6
5	3	33.3	6	7.0	40.3
6	3	35.3	5	6.7	42.0

Table 4.8: Case 7.

production runs of low-priority products are bounded in these environments.

4.1.4 Conclusions

The present section has presented an exact analysis of a two-queue state-dependent polling system with k-limited service extending the polling literature on both the k-limited service discipline and on state-dependent setups; containing the non-preemptive priority model and the model of Lee [145] as special cases. Moreover, the results of the analysis have been applied to a two-product case of the SELSP, which provides us with theoretical evidence that the k-limited strategy leads to considerable cost reductions compared to widely used (standard) exhaustive policies. The generating functions derived in the present section provide an excellent breeding ground for the development of simple (closed-form), accurate and efficient approximations for tail probabilities along the lines of the dominant pole approximation as described in, e.g., Tijms [208]. In that respect, our work may complement the results in recent work of Chang and Down [61; 62].

As stated in Chapter 1, the k-limited policy violates the branching property for polling systems implying that extensions of the analysis of the present section to more realistic systems are, in most likelihood, outside the borders of possibility and that one has to resort to approximations in these cases. In that sense, it is important to observe that one of the key steps of our approach (the reduction of the multi-dimensional stochastic process to a single dimension represented by the function $\alpha(\cdot)$) already breaks down in the simplest extension of the model, i.e., a two-queue case where both queues are served according to the k-limited service discipline. This observation unearthes the fact that the present study has indeed reached the borders of tractability. Therefore, in the following section we develop an efficient and accurate approximate decomposition approach for k-limited polling systems under the assumption of generally distributed arrival, service and setup distributions. The section is closed with a remark.

Remark 4.1.5 *[J.S.H. van Leeuwaarden, personal communication].* As mentioned at various places throughout the present monograph, polling systems find a variety of applications in many fields (in particular, k-limited polling systems have proved their merit in computer and communication systems [43; 63]). In the context of the present section, we want to mention the application in the control of traffic lights. In polling jargon, the stream that is being given green light corresponds to the queue receiving service. As explained in Van den Broek [56], the k-limited policy can be effectively applied to mimic the so-called fully-actuated control strategy for traffic lights, where the traffic lights are controlled based on the presence of vehicles. The analysis of the present section can be extended by assuming that vehicles arrive according to a general discrete process - note that the Poisson distribution is also a discrete distribution - and therewith model the system in discrete time. Such a discrete arrival process is the common assumption in traffic engineering (see, for example, Van den Broek *et al.* [57]) and allows us to study distributions with a larger coefficient of variation, distributions with a finite support or distributions fitted to empirical data. Van den Broek [56] states that in the past traffic engineers were sceptic about the fully-actuated control strategy, although they certainly recognized the practical potential, due to its reliance on heuristics rather than on exact analysis for dimensioning. Therefore, the exact analysis of the present section can be regarded as a valuable methodological contribution in the field of traffic engineering as well. □

4.2 Systems with multiple queues

The present section, which is an abridged version of [P8], aims to approximate the marginal queue length distributions in a continuous-time polling system with k-limited service under the assumption of general arrival, service and setup distributions. A feasible *approximate* approach for the queue length distribution in a k-limited polling system is the decomposition method, in which the polling system is decomposed in vacation systems, for which the vacation distributions are computed in an iterative approximate manner. At each step in the iteration the mathematical analysis focuses on a single queue, whereas the other queues in the system determine the length of the vacation period. This decomposition method is adopted by the present research as well. We have to remark that decomposition methods seem to be applicable to a wide variety of queueing systems (see, e.g., [77; 112; 216; 217]). In the past, some systems related to the one of the present section have been studied by the decomposition approach, i.e., a k-limited polling system with finite buffers under the assumption of Poisson arrival processes [139] or a k-limited polling system in combination with a reservation mechanism [146]. The qualitative observations of these studies seem to carry over to the system of the present section.

The key observation, which is at the same time the mathematical motivation of the present study, is the fact that it is extremely important to capture the correlations among the different queues, since these correlations have a significant impact on the performance measures. Whereas [139] does not take these dependencies into account, [146] proposes to take a weighted sum of a completely uncorrelated and a perfectly correlated system in each step of the iteration by using a pre-defined mixing probability. Although the method of [146] clearly outperforms the procedure that ignores the correlations, this procedure is unable to compensate for correlations in systems with only two queues and also is difficult to apply for systems with more than two queues. That is, since the quality of the procedure strongly depends on the mixing probability, it is rather complicated to find an expression for this probability providing accurate results over the entire range of parameters. Further, the procedure of [146] is based on generating functions, the numerical determination of zeros and the numerical inversion of characteristic functions, considerably increasing the computational complexity of the algorithm. Finally, due to special features of the protocol studied in [146] the correlations between the queue lengths are relatively small compared to our system (e.g., in case all queues have a service limit of 1 the correlations vanish), which makes the approach of [146] well suited for that particular protocol.

Therefore, the goal of the present study is the development of a computationally efficient iterative approximation method for the marginal queue length distributions in the k-limited polling model. The main challenge can be found in the estimation of the correlations between the queue lengths in each step of the iterative algorithm. The vast majority of the literature on polling systems is devoted to waiting time figures, while almost no attention has been given to the analysis of such correlations. Recall that in Chapter 3 we have derived heavy-traffic asymptotics for the covariances between successive station times in polling systems with mixtures of *gated* and *exhaustive* service under the assumption of Poisson arrivals. However, to the best of our knowledge no results are known for the correlations among queues in polling systems with *k-limited* service.

The key ideas of the approach undertaken in the present section for polling systems with k-limited service are as follows:

1. The dependence between the queue under consideration and the other queues is taken into account by the introduction of conditional vacations (also called intervisit periods), i.e., the length of the intervisit period is positively correlated to the length of the preceding visit period.

2. The mutual dependencies of the other queues are approximated via standard probabilistic arguments and the conditional intervisit periods.

The main contribution of the present section is the development of a novel iterative approximation scheme for k-limited polling systems with general arrival, service and setup distributions. The algorithm developed in the present section only needs information on the first two moments of all distributions. The accuracy of the approximation scheme is verified by means of an extensive simulation study. The approximation scheme turns out to be robust and computationally efficient, while the differences between the exact and approximate values are small within a reasonable margin. In particular, the time complexity is only polynomial in the number of queues and the service limits. The main building block of this algorithm is a k-limited service vacation model with state-dependent vacations, which has not been studied before in the open literature. In this vacation model, the vacation length depends on the length of the preceding visit period to the queue. As a spin-off, we present an exact analysis for this vacation model with the help of matrix-analytic techniques. A final word on the applicability of the algorithm is that it can also be used as approximation for the exhaustive discipline by taking a "large" value of the service limits. Therefore, our algorithm can also be seen as extension of the algorithm of Federgruen and Katalan [89] for the exhaustive polling system with Poisson arrivals to systems with general arrival processes.

The rest of the present section is organized as follows. Subsection 4.2.1 gives, besides the introduction of the model and further notation, a high-level view of the approximation scheme. In Subsection 4.2.2 the approximations for the mean and the variance of the conditional intervisit period are presented. Building on these results, Subsection 4.2.3 analyzes a k-limited vacation model with state-dependent vacations. Subsection 4.2.4 contains an overview of the iterative procedure to calculate the performance measures of interest. An extensive numerical study to test the accuracy of the approximation algorithm is presented in the penultimate subsection. Finally, the last subsection describes the main conclusions of the present research and indicates some possible directions for further research.

4.2.1 Model description

We consider the basic N-queue polling system described in Chapter 2, where each queue is served according to the k-limited policy. In particular, it is assumed that the setup times are *state-independent* again. However, the Poisson arrival processes introduced in Chapter 2 are extended implying that we assume that customers arrive at all queues according to independent (general) processes. The mean and second moment of the interarrival times are denoted by $\mathbb{E}[A_i]$ and $\mathbb{E}[A_i^2]$, $i = 1, 2, \ldots, N$, respectively. Our main interest is in L_i, the queue length (including the customer possibly in service) at queue i at an arbitrary point in time, $i = 1, 2, \ldots, N$. The main result of the present section is the development of an iterative scheme to approximate the *complete* distribution of L_i.

We continue with a high-level description of our approximation method. The key approximation idea is that we decompose the original k-limited polling system with N queues into a set of N separate k-limited *single*-queue models with vacations. At each step in the iteration the mathematical analysis focuses on a single queue i, whereas the other queues in the system determine the length of the vacation period (intervisit period) of queue i,

$i = 1, 2, \ldots, N$. The bottleneck in this approximation is the derivation of the distribution of the intervisit period, which will be done in an iterative way. If we assume that the distribution of the intervisit period is known in step n of the iteration, the distribution of the visit period in step $n + 1$ is derived by means of a queueing analysis for the k-limited single-queue model with vacations (see Subsection 4.2.3). In its turn, the latter distribution can be used to compute the distribution of the length of the intervisit period in step $n + 1$ (see Subsection 4.2.2).

Since it is more likely that a long (short) visit period is followed by a long (short) intervisit period, conditional intervisit periods are introduced. That is, the length of an intervisit period is assumed to be positively correlated to the number of customers served in the preceding visit period. The subsequent two subsections aim to answer the following questions:

1. What are the first two moments of an intervisit period for queue i given that $l = 0, 1, \ldots, k_i$ customers are served in queue i in the preceding visit period (see Subsection 4.2.2)?

2. What is the distribution of the number of customers served in a visit period for queue i given the first two moments of the conditional intervisit periods (see Subsection 4.2.3)?

4.2.2 Intervisit period

The present subsection computes the first two moments of an intervisit period for queue i given that $l = 0, 1, \ldots, k_i$ customers are served in queue i in the preceding visit period. The input of the present subsection are the stationary probabilities $\pi_j(l)$ that l customers are served during this visit period of queue j. These probabilities follow from the analysis of the vacation model in the previous iteration step as expounded in Subsection 4.2.3. For presentation reasons, we omit throughout this subsection the superscript n in all random variables denoting the corresponding iteration step n.

First moments. The intervisit period of a queue i is obviously positively correlated to the preceding visit period of queue i, $i = 1, 2, \ldots, N$. Therefore, we introduce so-called *conditional* visit periods $V_i(l)$, intervisit periods $I_i(l)$ and cycles $C_i(l)$ conditioned on the number of customers $D_i = l$ served in the visit period of queue i, $l = 0, 1, \ldots, k_i$.

The mean conditional cycle lengths may be approximated by using approximate balance equations for $C_i(l)$ as proposed by [138],

$$(\rho - \rho_i)\mathbb{E}[C_i(l)] + l\mathbb{E}[B_i] \approx \mathbb{E}[C_i(l)] - \mathbb{E}[S], \qquad i = 1, 2, \ldots, N, \quad l = 0, 1, \ldots, k_i, \quad (4.31)$$

which equate the amount of work arriving (left hand side) and the amount of work departing during conditional cycles (right hand side). The balance equation (4.31) is obviously an approximation, since it assumes balance within each conditional cycle which may not hold. Notice the similarity with the *exact* balance equation for the *unconditional* cycle length, for which work-in is equal to work-out. Solving (4.31) results in

$$\mathbb{E}[C_i(l)] \approx \frac{l \cdot \mathbb{E}[B_i] + \mathbb{E}[S]}{1 - \rho + \rho_i}, \qquad i = 1, 2, \ldots, N, \quad l = 0, 1, \ldots, k_i. \quad (4.32)$$

We extend the approximation of [138] by multiplying the individual values $\mathbb{E}[C_i(l)]$ with a scaling factor $c_i \in \mathbb{R}$,

$$\mathbb{E}[C_i(l)] \approx c_i \frac{l \cdot \mathbb{E}[B_i] + \mathbb{E}[S]}{1 - \rho + \rho_i}, \qquad i = 1, 2, \ldots, N, \quad l = 0, 1, \ldots, k_i, \quad (4.33)$$

and we set the c_i as

$$c_i = \frac{\mathbb{E}[C]}{\sum_{l=0}^{k_i} \pi_i(l)\mathbb{E}[C_i(l)]}, \qquad i = 1, 2, \ldots, N, \tag{4.34}$$

which implies that the correct unconditional cycle length as given by (2.5) is maintained. This scaling obviously facilitates the convergence and stability of the algorithm.

Then, the mean conditional intervisit periods $I_i(\cdot)$ can be approximated in the following way,

$$\mathbb{E}[I_i(l)] \approx \mathbb{E}[C_i(l)] - l \cdot \mathbb{E}[B_i], \qquad i = 1, 2, \ldots, N, \quad l = 0, 1, \ldots, k_i. \tag{4.35}$$

Finally, we define a conditional visit period $V_i^j(l)$ as the length of the visit period of queue j given that in the preceding visit to queue i precisely l customers are served, $l = 0, 1, \ldots, k_i$. The mean of this random variable reads

$$\begin{aligned} \mathbb{E}[V_i^j(l)] \ &\approx \ \rho_j \mathbb{E}[C_i(l)], \qquad && i = 1, 2, \ldots, N, \quad l = 0, 1, \ldots, k_i, \\ & && j = i+1, \ldots, N, 1, \ldots, i-1, \end{aligned} \tag{4.36}$$

which completes the analysis of the conditional first moments.

We have to remark that the approximations of the present subsection only compensate for the correlations between the visit period and the *immediately following* intervisit period. Although it is not inconceivable that one may come up with more sophisticated approximations, the numerical evaluation of Subsection 4.2.5 shows that our approximations are still very effective in capturing the correlations among the queues.

Second moments. The goal of the present subsection is the development of an approximation for the variance of the *conditional* intervisit periods $I_i(\cdot)$. The starting point of our analysis are the *unconditional* intervisit periods I_i. Since the setup times are assumed to be uncorrelated (see Chapter 2), the variance of such an unconditional intervisit period I_i is given by

$$\mathbb{V}\mathrm{ar}[I_i] = \sum_{j \neq i} \mathbb{V}\mathrm{ar}[V_j] + \sum_{j} \mathbb{V}\mathrm{ar}[S_j] + 2 \sum_{j \neq i} \sum_{\substack{k > j \\ k \neq i}} \mathbb{C}\mathrm{ov}[V_j, V_k] + \sum_{\substack{j \\ k \neq i}} \mathbb{C}\mathrm{ov}[S_j, V_k], \tag{4.37}$$

where the latter two summations include all the covariances among the various visit periods and among the setup times, respectively, within an intervisit period of queue i. Therefore, the $>$ sign in this summation means that queue k is visited after queue j in this intervisit period.

The terms $\mathbb{V}\mathrm{ar}[V_j]$ in the right-hand side of (4.37) represent the variance of unconditional visit periods V_j of queue j. The second moment of such a visit period can be approximated as follows. Conditioning on the number of customers served during the visit period of this queue and ignoring the correlations between the length of the service times and the number of customers served during the visit period yields

$$\mathbb{E}[V_i^2] = \sum_{l=0}^{k_i} \pi_i(l)\mathbb{E}[V_i^2(l)] \approx \sum_{l=0}^{k_i} \pi_i(l)(l\mathbb{E}[B_i^2] + l(l-1)\mathbb{E}[B_i]^2), \qquad i = 1, 2, \ldots, N. \tag{4.38}$$

Now, the variance of V_i can be obtained via standard probabilistic arguments.

Since the terms $\mathbb{V}\mathrm{ar}[S_j]$ are assumed to be input of the system, one does not need to approximate them. By definition, the covariance terms $\mathbb{C}\mathrm{ov}[V_j, V_k]$ appearing in (4.37) can be rewritten as

$$\mathbb{C}\mathrm{ov}[V_j, V_k] = \mathbb{E}[V_j V_k] - \mathbb{E}[V_j]\mathbb{E}[V_k], \tag{4.39}$$

where the terms $\mathbb{E}[V_j]$ and $\mathbb{E}[V_k]$ follow from (2.7). To compute the unknown quantity $\mathbb{E}[V_jV_k]$, we condition on the number D_j of customers served in queue j during the last visit period as follows

$$\begin{aligned}\mathbb{E}[V_jV_k] &= \sum_{l=0}^{k_j} \mathbb{E}[V_jV_k|D_j = l]\pi_j(l) \\ &\approx \sum_{l=0}^{k_j} l\mathbb{E}[B_j]\mathbb{E}[V_j^k(l)]\pi_j(l), \qquad (4.40)\end{aligned}$$

where $\pi_j(l)$ follow from the analysis of Subsection 4.2.3 and $\mathbb{E}[V_j^k(l)]$ can be approximated by (4.36).

Finally, in case a queue k is visited before queue j in the intervisit period of queue i, V_k and S_j are obviously uncorrelated. In case queue j is visited first, we assume independence between setup times and visit periods as well, i.e.,

$$\mathbb{C}\text{ov}[S_j, V_k] \approx 0, \qquad (4.41)$$

and, thus, all terms in (4.37) have been specified. Asymptotic numerical results in Van Vuuren and Winands [P8] show that this assumption is valid as long as the setup times are not too variable.

By definition, the coefficient of variation c_{I_i} of an unconditional intervisit period is, subsequently, given by

$$c_{I_i} = \frac{\sqrt{\mathbb{V}\text{ar}[I_i]}}{\mathbb{E}[I_i]}, \qquad i = 1, 2, \ldots, N. \qquad (4.42)$$

We approximate the variance of the conditional intervisit periods $I_i(\cdot)$ by assuming equality of the coefficients of variation of all periods, i.e.,

$$\mathbb{V}\text{ar}[I_i(l)] \approx c_{I_i}^2 \cdot \mathbb{E}[I_i(l)]^2, \qquad l = 1, 2, \ldots, k_i, \quad i = 1, 2, \ldots, N, \qquad (4.43)$$

where an approximation of $\mathbb{E}[I_i(\cdot)]$ is given by (4.35). We add that we have also experimented with other approximations for the variance of conditional visit period such as assuming equality of the coefficients of variation of all conditional cycle lengths. Approximation (4.43), however, turned out to be the most accurate one. Finally, notice that the expression in the righthand side of (4.43) is increasing in l.

4.2.3 Visit period

The present subsection aims to compute the distribution of a visit period for queue i given the first two moments of the conditional intervisit periods as computed via (4.35) and (4.43) in the preceding subsection. By means of matrix-analytic techniques, we analyze a single-station vacation model with k-limited service, in which the vacation length depends on the length of the preceding visit period. The author is aware of only one other study in which this specific dependency is studied (under the restrictive assumption of Poisson input [153]). Comprehensive surveys on vacation models can be found in [79; 80; 204].

Since the present subsection is focussing on a single queue i in a specific iteration step n, the subscript i and superscript n are dropped from all random variables. Throughout the present subsection, the distribution functions of the interarrival and the service times are needed. However, the only information available for these random variables are the first two moments. A common way to obtain an *approximate* distribution is to fit a phase-type distribution on the first two moments as elucidated in Van Vuuren and Winands [P8] (cf., e.g., [208]). In the remainder of the present subsection, we assume that the fitted distributions are used as substitute for the arrival and service distributions and that the number of phases needed equal n_A and n_B, respectively.

In the preceding subsection, we have computed the first two moments of the conditional intervisit periods $I(\cdot)$ conditioned on the exact number of customers served in the preceding visit period. To keep the size of the state space for the k-limited vacation model manageable, some of these intervisit periods are aggregated. That is, we draw a distinction between intervisit periods $I(0)$, $I(k)$ and $I(*)$ in which there have been zero, the maximum number or any other number of customers served in the preceding visit period, respectively. In case the service limit at a queue equals one, only $I(0)$ and $I(1)$ have to be distinguished. The period $I(*)$ is, thus, defined as,

$$I(*) := \frac{\sum_{l=1}^{k-1} \pi(l) I(l)}{\sum_{l=1}^{k-1} \pi(l)}, \tag{4.44}$$

with first two moments,

$$\mathbb{E}[I(*)] := \frac{\sum_{l=1}^{k-1} \pi(l)\mathbb{E}[I(l)]}{\sum_{l=1}^{k-1} \pi(l)}, \quad \text{and} \quad \mathbb{E}[I(*)^2] := \frac{\sum_{l=1}^{k-1} \pi(l)\mathbb{E}[I(l)^2]}{\sum_{l=1}^{k-1} \pi(l)}, \tag{4.45}$$

where the $\pi(l)$ follow from the previous iteration step. We have tested this aggregation of intervisit periods for a wide variety of cases, from which we concluded that it has only negligible (negative) impact on the results, which is outweighted by the gain in efficiency.

In sum, the system under consideration is a single-server k-limited vacation model with three different kinds of intervisit periods dependent on the number of customers served in the preceding visit period. In order to construct these intervisit periods in an efficient way, we introduce the auxiliary mutually independent random variables $\tilde{I}(*)$ and $\tilde{I}(k)$, which are independent of $I(0)$ as well. These random variables satisfy

$$I(*) = \tilde{I}(*) + I(0), \quad \text{and} \quad I(k) = \tilde{I}(k) + I(*), \tag{4.46}$$

which is always possible since the variances of the conditional intervisit periods are increasing in l as shown in (4.43). Thereupon, phase-type distributions are fitted on $I(0)$, $\tilde{I}(*)$ and $\tilde{I}(k)$ (see Van Vuuren and Winands [P8] for further details) in such a way that the first two moments of $I(*)$ and $I(k)$ are correct. If we assume that the number of phases needed for the description of $I(0)$, $\tilde{I}(*)$ and $\tilde{I}(k)$ equal $n_{I(0)}$, $n_{\tilde{I}(*)}$ and $n_{\tilde{I}(k)}$, respectively, the total number n_I of phases for the intervisit process is given by $n_I = n_{I(0)} + n_{\tilde{I}(*)} + n_{\tilde{I}(k)}$.

The k-limited vacation model can be described by a continuous-time Markov process with states (i, j, m). The state variable $i = 0, 1, \ldots$ denotes the total number of customers in the specific queue under consideration, whereas the state variable $j = 1, 2, \ldots, n_A$ indicates the phase of the arrival process A. Finally, $m = 1, 2, \ldots, n_D$ indicates the phase of the departure process D, which is the combination of the service process and vacation processes $I(0)$, $\tilde{I}(*)$ and $\tilde{I}(k)$. These latter two processes can be modeled by a single variable, since the server is either serving customers or is on vacation. When the server is serving customers, one has to keep track of the phase of the service process and of the number of customers already served in the corresponding visit period. On the other hand, when the server is on vacation the phase of the corresponding vacation period is needed. Consequently, the total number of states for the departure process is $n_D = k \times n_B + n_I$. The phases of this departure process are grouped as follows: first, we group all phases related to the k service processes and, then, the phases of $\tilde{I}(k)$, $\tilde{I}(*)$ and $I(0)$.

Refer by level i to the set of states with i customers in the system and group the states by these levels, so that (i, j, m) precedes (i', j', m') if $i < i'$. Within each level, the states are grouped according to the arrival phase, so that (i, j, m) precedes (i, j', m') if $j < j'$. Lastly, the states are ordered by the departure phase, so that (i, j, m) precedes (i, j, m') if $m < m'$. Now, one may verify that the introduced Markov process is a *quasi-birth-and-death* (QBD) process where the infinitesimal generator $\mathbf{Q}$ has the following block-tridiagonal structure,

$$\mathbf{Q} = \begin{pmatrix} \mathbf{B}_{00} & \mathbf{B}_{01} & & & \\ \mathbf{B}_{10} & \mathbf{A}_1 & \mathbf{A}_0 & & \\ & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.47}$$

Below we specify the submatrices in $\mathbf{Q}$, where we use the concept of *Markovian Arrival Process* (MAP) (see, e.g., [37]) to describe the arrival and departure processes. In general, a MAP is defined in terms of a continuous-time Markov process with finite state space $\{0, \cdots, m-1\}$ and generator $\mathbf{G}_0 + \mathbf{G}_1$. The element $\mathbf{G}_1(i, j)$ denotes the intensity of transitions from i to j accompanied by an arrival. For $i \neq j$ element $\mathbf{G}_0(i, j)$ denotes the intensity of the remaining transitions from i to j, while the diagonal elements $\mathbf{G}_0(i, i)$ are strictly negative and chosen such that the row sums of $\mathbf{G}_0 + \mathbf{G}_1$ are zero.

The arrival process can be straightforwardly represented by such a MAP, the states of which correspond to the phases of this process. Its generator can be expressed as $\mathbf{G}_0^A + \mathbf{G}_1^A$, where the transition rates in $\mathbf{G}_1^A$ are the ones that correspond to an arrival of a customer to the system. For the transition rates of the $\mathbf{G}_0^A$ and $\mathbf{G}_1^A$ matrices, we refer to Van Vuuren and Winands [P8].

The MAP for the departure process with generator $\mathbf{G}_0^D + \mathbf{G}_1^D$ is a little more involved. All transitions related to the vacation periods do not cause departures and are, thus, within $\mathbf{G}_0^D$. Completion of a service process, obviously, leads to a departure implying that the corresponding rates are in $\mathbf{G}_1^D$. Transitions within a service process not causing departures are, of course, part of $\mathbf{G}_0^D$. Further, we have to distinguish between the situation when there are two or more customers in the system or not. In the first situation, if a departure is not the k^{th} departure the next service process is started and if it is the k^{th} departure a new vacation period is begun. To deal with the situations in which there are only zero or one customers present, we have to introduce matrices $\tilde{\mathbf{G}}_0^D$ and $\tilde{\mathbf{G}}_1^D$, representing the transition within level 0 and the transitions from level 1 to level 0, respectively. We can recognize two differences between these matrices and $\mathbf{G}_0^D + \mathbf{G}_1^D$. First, when a service process is completed which is not the k^{th} service, a vacation period is commenced instead of the next service. Second, when a vacation period is finished, we jump to process $I(0)$ instead of to the service process of the first customer in the visit period. The transition rates for $\mathbf{G}_0^D$, $\mathbf{G}_1^D$, $\tilde{\mathbf{G}}_0^D$ and $\tilde{\mathbf{G}}_1^D$ are, again, summarized in Van Vuuren and Winands [P8].

Now, we are in the position to describe all the submatrices in $\mathbf{Q}$, i.e.,

$$\mathbf{B}_{01} = \mathbf{G}_1^A \otimes \mathbf{I}_{n_D}, \tag{4.48}$$

$$\mathbf{B}_{00} = \mathbf{G}_0^A \otimes \mathbf{I}_{n_D} + \mathbf{I}_{n_A} \otimes \tilde{\mathbf{G}}_0^D, \tag{4.49}$$

$$\mathbf{B}_{10} = \mathbf{I}_{n_A} \otimes \tilde{\mathbf{G}}_1^D, \tag{4.50}$$

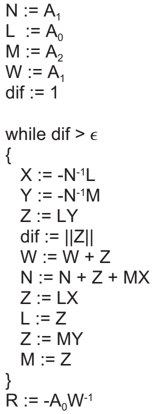
$$\mathbf{A}_0 = \mathbf{G}_1^A \otimes \mathbf{I}_{n_D}, \tag{4.51}$$

$$\mathbf{A}_1 = \mathbf{G}_0^A \otimes \mathbf{I}_{n_D} + \mathbf{I}_{n_A} \otimes \mathbf{G}_0^D, \tag{4.52}$$

$$\mathbf{A}_2 = \mathbf{I}_{n_A} \otimes \mathbf{G}_1^D, \tag{4.53}$$

where $\mathbf{I}_n$ is the identity matrix of size n and if $\mathbf{A}$ is an $n_1 \times n_2$ matrix and $\mathbf{B}$ an $n_3 \times n_4$ matrix the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is an $n_1 n_3 \times n_2 n_4$ matrix defined by

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} \mathbf{A}(1,1)\mathbf{B} & \cdots & \mathbf{A}(1,n_2)\mathbf{B} \\ \vdots & & \vdots \\ \mathbf{A}(n_1,1)\mathbf{B} & \cdots & \mathbf{A}(n_1,n_2)\mathbf{B} \end{pmatrix}. \tag{4.54}$$



```
N := $A_1$
L  := $A_0$
M := $A_2$
W := $A_1$
dif := 1

while dif > $\epsilon$
{
  X := $-N^{-1}L$
  Y := $-N^{-1}M$
  Z := LY
  dif := ||Z||
  W := W + Z
  N := N + Z + MX
  Z := LX
  L := Z
  Z := MY
  M := Z
}
R := $-A_0W^{-1}$
```

Figure 4.1: Algorithm of [169] for finding the rate matrix $\mathbf{R}$, where $\|.\|$ denotes a matrix-norm and ϵ some positive number.

This completes the description of the QBD. If we let $\mathbf{q}_i$ denote the equilibrium probability vector of level i, the corresponding balance equations are given by

$$\mathbf{q}_{n-1}\mathbf{A}_0 + \mathbf{q}_n\mathbf{A}_1 + \mathbf{q}_{n+1}\mathbf{A}_2 = 0, \quad n \geq 2, \tag{4.55}$$

and

$$\begin{aligned} \mathbf{q}_0\mathbf{B}_{00} + \mathbf{q}_1\mathbf{B}_{10} &= 0, & (4.56)\\ \mathbf{q}_0\mathbf{B}_{01} + \mathbf{q}_1\mathbf{A}_1 + \mathbf{q}_2\mathbf{A}_2 &= 0. & (4.57)\end{aligned}$$

Introducing the rate matrix $\mathbf{R}$ as the minimal nonnegative solution of the nonlinear matrix equation,

$$\mathbf{A}_0 + \mathbf{R}\mathbf{A}_1 + \mathbf{R}^2\mathbf{A}_2 = 0, \tag{4.58}$$

it can be proved that the equilibrium probabilities satisfy (see, e.g., [170]),

$$\mathbf{q}_{n+1} = \mathbf{q}_n\mathbf{R}, \quad n \geq 1. \tag{4.59}$$

To determine this matrix $\mathbf{R}$ we use the algorithm developed by [169] as listed in Figure 4.1. The vectors $\mathbf{q}_0$ and $\mathbf{q}_1$ follow from the boundary conditions (4.56), (4.57), and the normalization condition. This queue length distribution $\mathbf{q}_i$ yields the following expression for the distribution of the number of customers served in a visit period,

$$\pi(l) = \frac{h(l)}{\sum_{i=0}^{k} h(i)}, \qquad l = 0, 1, \ldots, k, \tag{4.60}$$

where $h(l)$ is the total rate of jumps to a vacation period after serving l customers. To calculate $h(l)$ we have to sum all transition rates from a state where $l-1$, $l = 1, 2, \ldots, k$, customers are served (or 0 customers when $l = 0$) to a vacation, multiplied by the probability of being in that specific state. Further, we recall that the indices of $\mathbf{q}_{\cdot}(\cdot)$ within the brackets correspond to lexicographically ordered states of the arrival and departure processes. So,

$$\begin{aligned} h(0) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_{I(0)}} \Big(\mathbf{q}_0((i-1)n_D + kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + j) \times \qquad (4.61) \\ &\quad \mathbf{B}_{00}((i-1)n_D + kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + j, (i-1)n_D + kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + 1)\Big), \\ h(l) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \mathbf{q}_1((i-1)n_D + (l-1)n_B + j) \times \\ &\quad \mathbf{B}_{10}((i-1)n_D + (l-1)n_B + j, (i-1)n_D + kn_B + n_{\tilde{I}(k)} + 1), \qquad (4.62) \\ &\qquad l = 1, \dots, k-1, \\ h(k) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \mathbf{r}((i-1)n_D + (k-1)n_B + j) \times \\ &\quad \mathbf{A}_2((i-1)n_D + (k-1)n_B + j, (i-1)n_D + kn_B + 1), \qquad (4.63) \end{aligned}$$

where

$$\mathbf{r} = \sum_{i=1}^{\infty} \mathbf{q}_i = \sum_{i=1}^{\infty} \mathbf{q}_1 \mathbf{R}^{i-1} = \mathbf{q}_1 (\mathbf{I}_{n_A \times n_D} - \mathbf{R})^{-1}, \tag{4.64}$$

which completes the analysis of the k-limited vacation model.

4.2.4 Iterative algorithm

As described at the end of Subsection 4.2.1, the performance characteristics of the k-limited polling system are approximated by an iterative scheme. The algorithm is as follows.

Outline of the algorithm.

- *Step 0*: Choose initial characteristics for all queues (for details see below).
- *Step 1*: For $i = 1$ to N, determine the first two moments of the conditional intervisit period $I_i(\cdot)$ for queue i from (4.35) and (4.43), respectively.
- *Step 2*: For $i = 1$ to N, determine the distribution of the number of customers served in the visit period V_i from (4.60).
- *Step 3*: Repeat Steps 1 and 2 until the characteristics for all queues have converged.
- *Step 4*: For $i = 1$ to N, compute the performance measures of interest for queue i.

Initialization. In Step 0 of the algorithm, we have to choose initial values for $\pi_i(l)$, $l = 0, 1, \dots, k_i$ and $i = 1, 2, \dots, N$. The assumption is made that all of these probabilities are zero except for $\pi_i(k_i)$, $i = 1, 2, \dots, N$. Notice that, via the approach developed in Subsection 4.2.2, the correct mean cycle lengths are obtained as computed by (2.5). We note that we have experimented with a large number of initial values, from which we concluded that the starting values of the algorithm have no, or at least negligible, impact on the results.

Test bed				
Parameter	**Notation**	**Value**		
		low	*medium*	*high*
Number of queues	N	2	5	10
Load	ρ	0.45	0.60	0.75
Service limit	k_i	1	5	10
SCV interarrival times	A_i	0.25	1	2
SCV service times	B_i	0.25	-	1
SCV setup time	S_i	0.25	-	1
Imbalance interarrival times	I_{A_i}	1:1	-	1:10
Imbalance service time	I_{B_i}	1:1	-	1:10
Ratio service and setup times	I_{B_i/S_i}	1:1	-	10:1
	Number of instances		2592	

Table 4.9: Test bed.

Convergence criterion. After Steps 1 and 2 we check whether the iterative algorithm has converged by comparing the probabilities $\pi_i(\cdot)$, $i = 1, 2, \ldots, N$, in the $(n-1)$-th and n-th step. We decide to stop when the maximum of the absolute values of the differences is less than ε; otherwise we repeat Step 1 and 2. Hence, the convergence criterion is

$$\max_{l=0,1,\ldots,k_i} \left| \pi_i^{(n)}(l) - \pi_i^{(n-1)}(l) \right| < \varepsilon, \qquad \forall_{i=1,2\ldots,N}, \tag{4.65}$$

where ε is chosen to be 10^{-4}. Of course, we may use other stop-criteria as well, e.g., mean queue lengths or mean intervisit periods.

Complexity analysis. The complexity of this method is as follows. Within the iterative algorithm, solving a subsystem, i.e., a vacation model, consumes most of the time. In a single iteration step N subsystems are solved. The number of iterations needed is difficult to predict, but in practice this number is about 10 to 15 iterations. The time consuming part of solving a subsystem is the calculation of the **R** matrix. This can be done in $O(n_i^3)$ time, where n_i is the size of the **R** matrix of subsystem i. Then, the time complexity of one iteration becomes $O(N \max_i(n_i^3))$. This means that the time complexity is polynomial in the number of queues, the service limits and the number of phases for each process.

4.2.5 Numerical validation

The present subsection reports on an extensive numerical study designed to assess the accuracy of the approximation method developed. We compare the first two moments and tail probabilities of the queue length distribution with the ones produced by discrete event simulation. Each simulation run is sufficiently long such that the widths of the 95% confidence intervals of the performance measures of interest are smaller than 1% of the predicted value. A first important remark is that the computation time of our algorithm is considerably less than the simulation time, which can mount up to fifteen minutes or more. This inefficiency of simulation techniques for (k-limited) polling systems has been observed before by, e.g., [41].

Parameter setting. We use a broad set of parameters for the tests. The number of queues in the system is varied between 2, 5 and 10, whereas the service limits are either 1, 5 or 10. The total load on the system varies between 0.45, 0.60 and 0.75; as mentioned in Chapter 2 this load does not include the setup times. Hence, especially for small values of the service limits k_i the effective load on the system is considerably higher. For this reason, some cases are unstable, meaning that (2.51) does not hold, and are thus removed from the test bed.

The squared coefficients of variation of the interarrival, service and setup times for each queue are identical and are varied between 0.25 and 2 and between 0.25 and 1, respectively. Since the variations in the setup and service times tend to be small in production systems - in contrast to telecommunication systems where heavy-tailed random variables are common - we only consider cases in which these variations are indeed relatively small. We refer the reader to [P8] for test cases with highly variable input. Furthermore, we test cases for which the setup times are 10 times smaller than the service times and cases for which setup and service times are equal.

Furthermore, both balanced and imbalanced polling systems are considered. In the balanced cases we set the arrival rates of all queues equal to 1. We test imbalance in the average interarrival times by making the load of the most heavily loaded queue 10 times higher than that of the least heavily loaded queue, and by letting the arrival rates of the other queues change linearly such that the overall mean arrival rate is maintained at 1. For example, in case of 5 queues we get arrival rates $(0.182, 0.591, 1.000, 1.409, 1.818)$. Testing imbalance in the service times proceeds along the same lines. This leads to a total of $3^4 2^5 = 2592$ test cases, which are summarized in Table 4.9. After removing the unstable cases, we end up with a total of 2088 cases. For further reference, we have classified the values for each parameter in the categories low, medium and high.

The performance measures under consideration in the present numerical study are the mean, standard deviation, 0.90-quantile and 0.95-quantile of the marginal queue length distributions, where the α-quantile of the distribution of a random variable X can be defined as the smallest value x such that

$$\mathbb{P}[X \leq x] \geq \alpha. \tag{4.66}$$

The importance of these quantiles lies in the fact that the optimal base-stock levels in the SELSP precisely equal these quantiles (see Chapter 1). More details on the standard deviation as being an important performance metric in production-inventory systems are given in Section 4.3.

Results. Table 4.10 summarizes the performance of the approach developed in the present section showing the average errors and for four error-ranges the percentage of the cases which fall in that range. Overall, we can say that for all performance measures the average error is around 7%, while the errors are for the majority of the cases less than 10%. We believe that these errors are in general satisfactory in view of the complexity of the system under consideration: we study a k-limited service discipline - containing the exhaustive policy as special case - under the assumption of general arrival processes, whilst the fact that our interest is in the complete queue length distribution constitutes an additional complicating factor.

To give this statement a more scientific basis, we compare the performance of our approach to the standard decomposition approach. In such a standard decomposition approach the dependencies among the individual queues are completely ignored. That is, the length of the intervisit period is assumed to be independent of the length of the preceding visit period, thus the need for conditional cycles and conditional (inter)visit periods cancels, and the correlations among the lengths of the individual visit periods are set equal to zero. Remark that the application of this standard approach to k-limited polling systems has not been published in the open literature.

The results for the latter approach are listed in Table 4.11. Comparing this table to Table 4.10, we can conclude that our approach not only halves the mean errors for all performance measures, but also that the standard approach, in contrast to our approach, quite

Errors approach of present section					
	Aver. (%)	0-10 %	10-20 %	20-30 %	> 30%
Mean queue lengths	7.26	76.25	17.77	5.12	0.86
SD queue lengths	8.34	71.02	20.16	5.51	3.30
0.90-quantile	6.58	75.62	14.80	5.75	3.83
0.95-quantile	7.33	73.37	15.95	6.80	3.88

Table 4.10: Overall results approach of present section.

Errors standard approach					
	Aver. (%)	0-10 %	10-20 %	20-30 %	> 30%
Mean queue lengths	15.40	40.95	30.94	15.61	12.50
SD queue lengths	15.26	40.37	29.98	16.91	12.74
0.90-quantile	13.45	57.95	16.52	11.69	13.84
0.95-quantile	13.26	54.02	17.10	14.08	14.80

Table 4.11: Overall results standard approach.

often results in more than 30% error. This observation clearly underpins the statement made in the introduction that it is extremely important to capture the correlations among the different queues, since these correlations have a significant impact on the performance measures. In particular, the performance of the standard approach significantly degrades as the total load increases as shown in Table 4.13, which is in agreement with our result in Chapter 3 that the correlation between successive station times converges to one as the total load tends to one for the cases of exhaustive and gated polling systems with Poisson arrivals. Table 4.12 shows that the accuracy of our approach decreases in heavy traffic as well; the decrease in accuracy is, however, not so severe as for the standard decomposition approach.

It would also be interesting to compare the performance of our approach to the one of the alternative approach developed in [146]. In this study, it is proposed to take a weighted sum of a completely uncorrelated and a perfectly correlated system in order to capture the correlations among the queues. A good choice of the desired mixing probability is an interesting problem in itself and the probability used in [146] has not been developed for the k-limited polling system covered in the present section, rather for a modification of this system, i.e., inclusion of a reservation mechanism. Directly applying the same mixing probability to our setting would certainly wrong the approach of [146] leading to an unfair comparison. Essentially, this observation reveals a weakness of the procedure of [146]: the quality of this procedure strongly depends on the choice of the mixing probability. Taking

Errors approach present section (%)			
	low	*medium*	*high*
Mean	4.43	6.72	11.64
SD	5.20	6.23	14.95
0.90-quantile	4.11	5.87	10.67
0.95-quantile	4.63	6.50	11.85

Table 4.12: Average errors for approach present section as function of ρ.

Errors standard approach (%)			
	low	*medium*	*high*
Mean	8.22	14.54	25.88
SD	8.32	14.09	25.78
0.90-quantile	10.11	9.22	22.75
0.95-quantile	6.06	11.71	25.55

Table 4.13: Average errors for standard approach as function of ρ.

Errors mean queue lengths (%)

Parameter	*low*	*medium*	*high*
N	8.96	7.17	5.74
ρ	4.43	6.72	11.64
k_i	9.35	6.91	6.39
A_i	6.70	6.96	8.14
B_i	6.79	-	7.74
S_i	6.92	-	7.61
I_{A_i}	7.32	-	7.19
I_{B_i}	5.17	-	9.51
I_{B_i/S_i}	5.07	-	8.67

Table 4.14: Average errors for the mean.

Errors SD queue lengths (%)

Parameter	*low*	*medium*	*high*
N	8.77	10.21	6.16
ρ	5.20	6.23	14.95
k_i	9.39	7.87	8.18
A_i	6.56	8.25	10.22
B_i	7.72	-	8.97
S_i	8.18	-	8.51
I_{A_i}	8.21	-	8.51
I_{B_i}	6.07	-	10.78
I_{B_i/S_i}	5.65	-	10.07

Table 4.15: Average errors for the SD.

Errors 0.90**-quantile** (%)

Parameter	*low*	*medium*	*high*
N	9.50	5.87	4.49
ρ	4.11	5.87	10.67
k_i	8.55	6.43	5.60
A_i	6.65	5.79	7.31
B_i	6.15	-	7.02
S_i	6.47	-	6.69
I_{A_i}	6.84	-	6.26
I_{B_i}	4.63	-	8.67
I_{B_i/S_i}	5.23	-	7.45

Table 4.16: Average errors for 0.90-quantile.

Errors 0.95**-quantile** (%)

Parameter	*low*	*medium*	*high*
N	7.61	9.23	5.25
ρ	4.63	6.50	11.85
k_i	9.29	6.90	6.60
A_i	6.59	7.51	7.87
B_i	7.02	-	7.64
S_i	7.08	-	7.57
I_{A_i}	7.67	-	6.90
I_{B_i}	5.04	-	9.78
I_{B_i/S_i}	5.85	-	8.27

Table 4.17: Average errors for 0.95-quantile.

the above into account, we confine ourselves to a more qualitative comparison between the two approaches. That is, when comparing the errors reported in [146] to the ones listed in Table 4.10, one can conclude that they are of the same order of magnitude. The approximation method of [146] has, however, only been tested in a system with smaller inherent dependencies for the special case of Poisson arrivals. We have to remark that Tables 4.14 through 4.17 show that the interarrival distribution has no or at least negligible effect on the accuracy of our approach.

More specifically, Tables 4.14 through 4.17 show the detailed results for our approach, when fixing one parameter at a certain level. When a row is partially empty, it means that this parameter is only tested on two levels. Our approximation method seems to be fairly insensitive to different parameter settings. In this respect, the parameter having the largest impact on the performance is the total utilization ρ as earlier illustrated in Table 4.12. Moreover, we observe that imbalance in the service times and a decrease in the setup times have negative impact on the accuracy, whereas the accuracy of our approach increases as the service limits become larger. This latter observation tempts one to use the approach of the present section as approximation for the exhaustive policy as well, as touched upon in Subsection 4.2.6.

Finally, the present subsection has shown the accuracy of the developed approximation for a wide range of cases. The test bed is, undoubtedly, not only representative for practical instances of the production application motivating the present research but also for most applications in communication systems. In Van Vuuren and Winands [P8], the applicability of the approximation is, however, evaluated beyond all limits and the accuracy of the approximation is tested thoroughly in the following *asymptotic* regimes:

1. Highly variable setup and/or service times;
2. Heavy traffic, i.e., $\rho \uparrow 1$;
3. Large setup times, i.e., $\mathbb{E}[S] \to \infty$;
4. Large number of queues, i.e., $N \to \infty$.

Due to the fact that polling systems, typically, show aberrant behavior in these asymptotic cases (see, also, Chapter 3), one cannot expect to be able to develop a single algorithm which is accurate both in standard traffic settings and for all possible asymptotic regimes. In Van Vuuren and Winands [P8] it is shown that the approach of the present section loses accuracy in the first two regimes, but compared to the standard approach still wins by a mile, whereas it becomes even more, or at least remains, accurate for the latter two regimes. From a practical point of view, the fact that our approach is less accurate in the first two asymptotic regimes is not a serious limitation since these situations only occur in ill-designed systems. That is, in case of highly variable input whencesoever, managers should obviously not rely on the k-limited but on the time-limited policy, while heavy traffic indicates improper dimensioning of the service limits. For more details on the results in the above asymptotic regimes, we refer to Van Vuuren and Winands [P8]. We close this subsection with a remark.

Remark 4.2.1 In the past, pseudo-conservation laws have been applied quite often to develop approximations for mean waiting times in polling systems (and, thus, mean queue lengths as well). Throughout the present section, we have deliberately left this approach aside, because our approach does not use this technique and because this technique only gives approximations for mean performance measures for the special case of Poisson arrivals (for more information see Chapter 2). An additional complexity that shows up when applying pseudo-conservations laws to polling systems with k-limited service is that in such systems these laws still contain some unknown terms that have to be approximated as shown in Chapter 2. Note that the most accurate algorithm [60] based on such a pseudo-conservation law can still give up to 20% errors for the mean waiting times in k-limited polling systems. □

4.2.6 Conclusions

In the present section, we have created a novel iterative approximation scheme for k-limited polling systems with general arrival, service and setup distributions to compute the complete queue length distributions. The multi-queue polling system has been decomposed into single-queue vacation systems with state-dependent vacations and k-limited service. We have analyzed this vacation model by means of matrix-analytic techniques under the assumption of general arrival, service and vacation processes. The main challenge was found in the computation of the correlations among the queues in each step of the iterative scheme. The accuracy of the approximation scheme has been validated via an extensive simulation study. The developed approximation turned out be accurate, robust and computationally efficient. As shown in Van Vuuren and Winands [P8], possible improvement of the algorithm may be obtained in heavy traffic and in cases with highly variable input. The numerical evaluation has shown that the algorithm converged relatively fast; a rigorous proof of convergence is, however, left as subject of further research.

With minor adjustments, the algorithm developed can be carried over to variants of the considered polling systems, e.g., systems with batch arrivals, discrete-time polling systems or systems with finite buffers. The latter extension is of practical interest, since it allows to evaluate the lost sales variant of the SELSP. Application of our algorithm to polling systems with so-called gated-type k-limited service, i.e., the servers serves only k customers in a queue who arrived before the server's visit, is also not inconceivable. A related remark is that for deterministic service times the k-limited coincides with the time-limited strategy with fixed time limits. By choosing service times with a negligible coefficient of variation as input, the algorithm of the present section can also be used for the evaluation of this time-limited policy. Moreover, due to the efficiency of the algorithm, it could be used directly as approximation for the standard exhaustive and gated policy as well by choosing a 'large' value for the service limits. In that sense, our algorithm may be considered as extension of the procedure of [89] for exhaustive and gated polling systems, which relies on a Poisson assumption. Finally, the algorithm of the present section may be extended to the computation of *derivatives* of performance measures with respect to the service limits. Such an extension would allow application of gradient methods to optimize system performance and sensitivity analysis with respect to these control variables. Due to the low computational complexity of the developed procedure, it can be used as subroutine in such an optimization procedure.

The present and preceding section have transformed the evaluation of the k-limited policy into a well-studied topic. The next section makes a step backwards and provides an explorative investigation of the quality of the k-limited policy itself.

4.3 Numerical evaluation

The goal of the present section, which stems from some unpublished material of the author, is to probe the exhaustive and k-limited policies. First of all, we compare the performance of both policies and discuss the value and behavior of optimal choices of the service limits (see Subsection 4.3.1). Second, we discuss the inclusion of idle times in the cycle, which has the potential to improve system performance (see Subsection 4.3.2).

Although it is a common belief that bounding visit times is an excellent tool for prioritizing among queues in polling systems, studies on the impact on overall system performance are scarce. An exception is [63], which is, however, mainly concerned with (overload) situations less relevant in the context of production applications. Triggered by this lacuna in the literature, we present an explorative study of the quality of the k-limited policy as means to prioritize among queues for improving total system performance. It is shown that the k-limited policy can significantly improve system performance - especially in asymmetric systems - proving the relevance of the quantity-limited lot-sizing policy in multi-item production settings.

Since it is impossible to find a single measure as a proxy for system performance, we study both the mean and standard deviation of the marginal queue length distribution. That is, our objectives are to find the service limits k_i that minimize the following weighted sums,

$$\min_{k_1,k_2,\ldots,k_N} \sum_{i=1}^{N} c_i \mathbb{E}[L_i] \qquad \text{and} \qquad \min_{k_1,k_2,\ldots,k_N} \sum_{i=1}^{N} c_i \sigma_{L_i}, \tag{4.67}$$

where c_i is the cost parameter of queue i, $i = 1, 2, \ldots, N$.

The importance of the mean queue length has been discussed at length in the present monograph, the importance of the standard deviation might need some explanation. First of all, within the cost-optimization framework introduced in Chapter 1, Zipkin [225] argues that the standard deviation captures the gross behavior of performance over a wide range of systems - as long as the input is not pushed to extremes - whereas no other performance metric of comparable simplicity does so. In a similar vein, it is shown that other important performance measures, such as the service level, are (nearly) proportional to the standard deviation (see, also, Gallego and Moon [109]). Furthermore, standard deviations can be used for estimating tail probabilities via Chebyshev's inequality [183] or for approximating the entire probability distribution via a two-moment fit [208]. Moreover, the standard deviation has to a lesser degree the disadvantage, opposed to more advanced cost functions, that extremely large effort is required in order to estimate the performance measure accurately, i.e., with small relative error or a narrow confidence interval. Finally, we should recall that the results of the present study are certainly not limited to the described production setting, but may be used in the design and optimization phase of many other fields of applications. For instance, in telecommunication systems one often wants to guarantee a homogeneous quality-of-service level expressed in the standard deviation of the queue length. Therefore, we refrain from testing a cost function only applicable for production environments.

4.3.1 Performance

The goal of the present subsection is not to test a myriad of cases, rather we analyze a number of representative cases in order to examine when the k-limited policy outperforms the exhaustive policy and the other way around. Within this examination, we compare the exhaustive policy with the k-limited policy for an optimal choice of the service limits. Optimization of k-limited systems is, however, even more intricate than evaluation implying that hardly any optimization studies have appeared in the vast polling literature. Borst *et al.* [43] develop an approximate approach to determine these service limits so as to minimize a weighted sum of the *mean* queue lengths, which has shown to be very effective. Van der Mei [160] presents extensions of the so-called power series algorithm, e.g., for systems with Bernoulli service, to calculate derivatives of performance measures which could be used in gradient methods for optimization. Similar procedures may be developed for the k-limited policy.

Due to this paucity of optimization results for k-limited polling systems for the standard deviation (and other performance measures), we resort to simple enumeration techniques for the optimization which causes no problems due to the size of our test bed. Recall, however, that the approximate algorithm developed in Section 4.2 possesses all properties required for the application of efficient and accurate gradient methods to optimize system performance. Since we do not want to be diverted by other effects induced, e.g., by inaccuracies of approximations, a discrete event simulation is used in the experiments instead of the approximation of the previous section. Each simulation run is sufficiently long such that the widths of the 95% confidence intervals are smaller than 0.25% of the predicted value.

Numerical results are presented in Tables 4.18 - 4.21, in which we compare the costs Z as defined in (4.67) for the optimal values of the service limits and the costs in case the exhaustive discipline is implemented for each queue. In these tables we restrict attention to deterministic setup time and service time distributions motivated by the observed robustness of the optimal service limits with respect to these distributions by Borst *et al.* [43]. The most important observation from these tables is that the k-limited policy outperforms the exhaustive policy in asymmetric systems due to either cost or physical factors (as concluded several times before in the present monograph).

As said, optimization of the service limits is a very intricate problem. In case of a symmetric system with setup times, we can however analytically prove that the *mean* queue lengths are minimized by setting the service limits at infinity. That is, in Levy *et al.* [152] it is proven for a wide class of policies that the total amount of unfinished work at time t in the system when the exhaustive policy is employed is smaller than when another policy is employed. An immediate consequence is that the steady-state mean amount of unfinished work in the system is minimized for the exhaustive policy, from which it follows that the mean queue lengths in symmetric systems are minimized by serving all queues

$N = 2$, $\rho = 0.6$, $\lambda_1 = \lambda_2$,
$\mathbb{E}[B_1] = \mathbb{E}[B_2] = 1$,
$\mathbb{E}[S_1] = \mathbb{E}[S_2] = 0.25$

	Mean		SD	
(c_1, c_2)	(k_1, k_2)	Z	(k_1, k_2)	Z
$(1, 1)$	$(\infty, 10)$	1.31	$(8, 8)$	1.82
	(∞, ∞)	1.31	(∞, ∞)	1.82
$(10, 1)$	$(\infty, 1)$	6.60	$(\infty, 2)$	9.11
	(∞, ∞)	7.22	(∞, ∞)	10.01
$(1, 10)$	$(1, \infty)$	6.60	$(1, \infty)$	9.11
	(∞, ∞)	7.22	(∞, ∞)	10.01

Table 4.18: Output system 1.

$N = 2$, $\rho = 0.6$, $\lambda_1 = 2\lambda_2$,
$\mathbb{E}[B_1] = \mathbb{E}[B_2] = 1$,
$\mathbb{E}[S_1] = \mathbb{E}[S_2] = 0.25$

	Mean		SD	
(c_1, c_2)	(k_1, k_2)	Z	(k_1, k_2)	Z
$(1, 1)$	$(\infty, 8)$	1.30	$(\infty, 8)$	1.79
	(∞, ∞)	1.30	(∞, ∞)	1.79
$(10, 1)$	$(\infty, 1)$	8.09	$(\infty, 1)$	10.13
	(∞, ∞)	8.63	(∞, ∞)	10.79
$(1, 10)$	$(2, 8)$	5.01	$(2, 8)$	7.79
	(∞, ∞)	5.67	(∞, ∞)	8.89

Table 4.19: Output system 2.

$N = 2$, $\rho = 0.6$, $\lambda_1 = 4\lambda_2$,
$\mathbb{E}[B_1] = \mathbb{E}[B_2] = 1$,
$\mathbb{E}[S_1] = \mathbb{E}[S_2] = 0.25$

	Mean		SD	
(c_1, c_2)	(k_1, k_2)	Z	(k_1, k_2)	Z
$(1, 1)$	(∞, ∞)	1.27	(∞, ∞)	1.71
	(∞, ∞)	1.27	(∞, ∞)	1.71
$(10, 1)$	$(\infty, 1)$	9.48	$(\infty, 1)$	11.05
	(∞, ∞)	9.78	(∞, ∞)	11.39
$(1, 10)$	$(2, 9)$	3.70	$(1, 9)$	6.49
	(∞, ∞)	4.20	(∞, ∞)	7.47

Table 4.20: Output system 3.

$N = 5$, $\rho = 0.75$, $\lambda_1 = \ldots = \lambda_2$,
$\mathbb{E}[B_1] = \ldots = \mathbb{E}[B_5] = 1$,
$\mathbb{E}[S_1] = \ldots = \mathbb{E}[S_5] = 0.05$

	Mean		SD	
(c_1, c_{2-5})	(k_1, k_{2-5})	Z	(k_1, k_{2-5})	Z
$(1, 1)$	(∞, ∞)	2.19	$(5, 6)$	3.78
	(∞, ∞)	2.19	(∞, ∞)	3.78
$(10, 1)$	$(7, 1)$	5.36	$(7, 1)$	9.45
	(∞, ∞)	6.14	(∞, ∞)	10.59
$(1, 10)$	$(1, 9)$	16.10	$(1, 9)$	28.24
	(∞, ∞)	17.96	(∞, ∞)	31.00

Table 4.21: Output system 4.

exhaustively. In case of asymmetric systems the numerics confirm the conjecture of [43] that if $c_i/\mathbb{E}[B_i]{=}\max_{j=1,2,\ldots,N} c_j/\mathbb{E}[B_j]$ then the service limit k_i should be equal to infinity (the cases where this conjecture does not seem to hold are most likely due to numerical inaccuracies of the simulation and the flat behavior of the objective function).

If we want to minimize the standard deviations, however, we are unable to pronounce upon the optimal values of the service limits, but what we do see is that the optimal service limits in this case are (almost) identical to the optimal service limits in case one wants to minimize the means. This suggests - although we realize that the tests have only been performed under a limited variety of environmental settings - that the approaches developed by [43] may also be directly applicable to minimization of standard deviations and, possibly, of other performance metrics as well. We want to stress that we have performed more tests, all confirming this observation. We close this subsection with a remark.

Remark 4.3.1 Although we have confined ourselves in the experiments to systems with a small or moderate number of queues, we can, however, argue what happens in the limit of an increasing number of queues (cf. Borst *et al.* [43]). That is, we can distinguish four (stable) cases as $N \to \infty$ for $i = 1, 2, \ldots, N$,

1. λ_i fixed, $\mathbb{E}[B_i] = O(1/N)$ and $\mathbb{E}[S_i]$ fixed;

2. $\lambda_i = O(1/N)$, $\mathbb{E}[B_i]$ fixed and $\mathbb{E}[S_i] = O(1/N)$;

3. λ_i fixed, $\mathbb{E}[B_i] = O(1/N)$ and $\mathbb{E}[S_i] = O(1/N)$;
4. $\lambda_i = O(1/N)$, $\mathbb{E}[B_i]$ fixed and $\mathbb{E}[S_i]$ fixed.

Since the mean number of customers arriving at queue i per cycle, $\lambda_i \frac{\mathbb{E}[S]}{1-\rho}$, approaches infinity in Case 1, each queue should be served exhaustively. In Case 2, the system reduces to a continuous polling system - where no distinction of different service disciplines exists - and, therefore, the choice of the service limits of k is irrelevant. In Cases 3 and 4, which are equivalent up to a scaling of time by N, taking $k_i = \infty$ for all i is often a good rule of thumb, since an increment of k_i by one typically reduces $c_i\mathbb{E}[L_i]$ and $c_i\sigma_{L_i}$ more than it increases $\sum_{j\neq i}^N c_j\mathbb{E}[L_j]$ and $\sum_{j\neq i}^N c_j\sigma_{L_j}$, respectively. Via rough reasoning this statement is made plausible in [43] for the mean queue lengths. □

4.3.2 Idle times

Fifteen years ago Sarkar and Zangwill [188; 221] published two papers, which caused a great deal of controversy. In the first paper, they numerically show that reduction of setup times can, counterintuitively, increase the mean queue lengths in polling systems and cyclic production systems (see, also, Example 4.3.2). In a subsequent provocative paper, Zangwill [221] contends that these results expose a flaw in Japanese production theory which is based on the benefits of reducing setup times. This latter paper triggered off a heated discussion in Interfaces [82; 111; 159; 222], which was accompanied by a statement of the editor-in-chief that [221] *alone drew more response than all other articles combined* while he was editor and we can add to this that the debate is still not closed.

Example 4.3.2 To illustrate the anomalous effect that insertion of idle times decreases the expected waiting time, we follow the analysis of [74]. That is, we consider a simple symmetric exhaustive polling systems for which the mean waiting times are given by (see Chapter 3),

$$\mathbb{E}[W] = \frac{\rho}{1-\rho}\mathbb{E}[R_B] + \frac{\mathbb{V}\text{ar}[S]}{2\mathbb{E}[S]} + \frac{\mathbb{E}[S]}{2}\frac{1-\rho/N}{1-\rho}. \tag{4.68}$$

If we now increase the total setup time in a cycle by a constant δ while the variance of the total setup time remains unchanged, (4.68) changes accordingly into

$$\mathbb{E}[W] = \frac{\rho}{1-\rho}\mathbb{E}[R_B] + \frac{\mathbb{V}\text{ar}[S]}{2\mathbb{E}[S]+2\delta} + \frac{\mathbb{E}[S]+\delta}{2}\frac{1-\rho/N}{1-\rho}, \tag{4.69}$$

where we remark that the first term is the standard $M/G/1$ term independent of δ and that the second and third term are decreasing and increasing in δ, respectively. From (4.69) we can derive the optimal shift δ^*,

$$\delta^* = \left(\sqrt{\mathbb{V}\text{ar}[S]\frac{1-\rho}{1-\rho/N}} - \mathbb{E}[S]\right)^+. \tag{4.70}$$

Now, we clearly see that the counterintuitive effect that the mean waiting times can be decreased by increasing the setup times occurs if

$$\frac{\mathbb{V}\text{ar}[S]}{\mathbb{E}[S]^2} > \frac{1-\rho/N}{1-\rho}. \tag{4.71}$$

This example shows that the persistent belief that reduction in setup times always leads to reduced waiting times, and thus queue lengths, is in general not true. □

As a side issue we want to mention - although we certainly do not want to intervene in the discussion on applicability and validity of [188; 221; 222] - that the observed counter-intuitive phenomenon only occurs if the setup times are "variable enough" (see Example 4.3.2 and [74]). As seen in (4.71) this requires unrealistically - at least for production settings - high coefficients of variation for these setup times, which tempts us to the statement that the significance of the observed effect should not be overestimated.

Motivated by the above observations, Federgruen and Katalan [91] study the impact of insertion of idle times on the queue length distribution and discover improvements in terms of cost reductions (even in systems without setup costs). In an accompanying paper [90], they prove that, for exhaustive and gated polling systems, a *single* idle time inserted prior to any of the queues can be used without loss of optimality implying that the search for optimal idle times can be reduced to that for a single scalar. Further, this implies that the queue length distribution is invariant with respect to deterministic shifts in setup times so long as the net shift is zero, *ceteris paribus*.

The main contribution of the present subsection is the extension of the result of [90] to a much more general setting. That is, we rigorously prove for a very general class of policies (including the exhaustive, gated and k-limited policies as special cases) that, if setup times are decreased (increased) by a fixed total amount, it is immaterial which specific setup times are decreased (increased). The only restriction enforced on the policy is that it is allowed to use local information only. Finally, for more information on the impact of setup times on (exhaustive and gated) polling systems we refer to [74] and the references therein.

Consider the basic polling system as introduced in Chapter 2. For the time being, we assume that the server follows the *exhaustive* service discipline, but this is extended later on. Let $S_i^{b_i} = S_i + b_i$ denote the setup time which results when the original setup time S_i is shifted by $b_i \in [-b_i^{max}, \infty]$ time units where $b_i^{max} \in \mathbb{R}^+$, with $\mathbb{R}^+$ the set of nonnegative reals, is the infimum of the support of S_i, $i = 1, 2, \dots, N$. Below, we show that the queue length distributions depend on the vector of shifts only via its aggregate sum. For the sake of presentation, we assume that the setup times themselves are deterministic (but this is not required for the result to hold as long as the shifts are deterministic). Without loss of generality, we focus on queue 1.

Theorem 4.3.3 *For a system with setup times $S_i^{b_i}$, the distribution of the queue length L_1 depends on the vector $(b_1, b_2, \dots, b_N)$ only via its sum b.*

Proof. We start with characterizing the impact of the shifts on the length of queue 1 at a polling instant of this queue. For that purpose, we consider the customers in the system to belong to one of the following two disjoint sets:

1. type-i customers arrived during setup times after the last visit to queue i, $i = 1, 2, \dots, N$;

2. all other customers.

First of all, we observe that the numbers of customers in both sets are obviously independent.

Moreover, at a polling instant of queue 1 the type-1 customers *in the first set* are those who arrived during the total setup time $\mathbb{E}[S] + b$ in the cycle. This number is a Poisson distributed random variable, which clearly shows that this subpopulation depends on the shifts only through their aggregate sum. Note that the numbers of type-i customers (with $i \neq 1$) in the first set are, however, dependent on the placements of the shifts.

Furthermore, the subpopulations of all customer types *in the second set* also are independent of the specific placement of the shifts. To see this, consider two identical systems which differ only in the placement of the shifts (but have the same total amount of shifts). Assume that at time 0 the numbers of customers of all subpopulations in the second set are

Input	
Queues	$N = 2$
Arrival	$(\lambda_1, \lambda_2) = (1.0, 1.0)$
Service	$(\mathbb{E}[B_1], \mathbb{E}[B_2]) = (0.25, 0.25)$
	$(c^2_{B_1}, c^2_{B_2}) = (1.0, 1.0)$
Setup	$(\mathbb{E}[S_1], \mathbb{E}[S_2]) = (0.25, 0.25)$
	$(c^2_{S_1}, c^2_{S_2}) = (5.0, 5.0)$
Limits	$(k_1, k_2) = (20, 20)$
Case 1	$b = 0.0$
Case 2	$b = 0.1$
Case 3	$b = 0.2$
Case 4	$b = 0.3$
Case 5	$b = 0.4$

Table 4.22: Input of the system.

identically distributed in both systems. For specificity, assume that time 0 was a polling instant of queue 1. This means that at time 0 the numbers of type-1 customers in *the first set* also are equally distributed, but the distributions of the number of type-i customers, where $i \neq 1$, in the first set may differ.

Hence, the *total* numbers of type-1 customers are identically distributed in both systems and, thus, the immediately following visit times of queue 1 in both systems are identically distributed. This establishes that at the next polling instant of the server (at queue 2), the numbers of customers of all subpopulations in the second set still follow the same (joint) distribution. Continuing this inductive argument shows that from time 0 on the subpopulations of all customer types *in the second set* will evolve in exactly the same way in both systems. This observation implies that, regardless of the specific placement of the shifts, also the cycles evolve according to the same probabilistic rules in both systems.

This leads to the conclusion that *at a polling instant* of queue 1 the total number of type-1 customers depends on the vector $(b_1, b_2, \ldots, b_N)$ only via its sum b. Since we have observed that the cycles are invariant under the placement of the shifts as long as the total shift remains constant, this implies that also *at arbitrary moments in time* the queue length distribution of queue 1 is dependent only on the sum of the shifts. □

Up to now, we have been focussing on the exhaustive discipline. However, the result presented holds for much more general service disciplines. That is, the result remains valid as long as the service discipline decides on the number of customers to serve only based on the number of customers present at the visited queue at the polling instant plus those who arrived during the visit time and not on the number of customers present and arriving at the other queues. For example, this means that also for the gated and the k-limited policy the result holds. Notice that this distinction has its origin in the fact whether the discipline incorporates global information or not.

We want to conclude the present subsection with an instance of a k-limited polling system for which including idle times actually improves performance (see Table 4.22 for the input). The output summarized in Table 4.23 shows that both the mean and standard deviation of the queue length can be slightly decreased by adding idle times, implying that the anomaly studied is not an idiosyncrasy of exhaustive and gated polling systems only, but can also occur in k-limited systems. The operational implication is that the developed quantity-limited policy can be straightforwardly extended with a single parameter, i.e., the inserted idle time in a cycle, which can lead to cost reductions and can be optimized by a standard method for minimization of nonlinear functions of a single variable.

Output				
	Mean	SD	0.90-q.	0.95-q.
Case 1	1.50	1.93	4.00	5.00
Case 2	1.47	1.87	4.00	5.00
Case 3	1.52	1.86	4.00	5.00
Case 4	1.61	1.88	4.00	5.00
Case 5	1.72	1.93	4.00	5.00

Table 4.23: Output of the system.

These results raise an interesting research direction (which we leave for further research), i.e., a large-scale numerical study examining not only the effect of idle times on the performance of general polling systems but also on the optimal values of the service limits. In this context, it is important to observe that inclusion of (deterministic) idle times and bounding the visit times have the same effect on the cycles, i.e., reduction of the variability in the cycle lengths. This subsection is closed with a remark.

Remark 4.3.4 For the exhaustive policy, [90] proves that the queue lengths are increasing in the setup times in the $\leq_m$ ordering (for the gated discipline this monotonicity is only shown for the first two moments). In Altman *et al.* [33] it is, however, shown that the queue lengths at polling instants are stochastically increasing in the sense of strong stochastic $\leq_{st}$ ordering - which is stronger than the $\leq_m$ ordering - for a very broad class of policies (of which branching-type and k-limited policies are members). Besides providing a strong foundation for the benefit of eliminating variability in setup times, this result shows that no gain can be achieved by inserting random instead of deterministic idle times. □

4.4 Conclusions

In the present chapter, we have analyzed the quantity-limited lot-sizing policy. Significant cost reductions have been observed by implementing the quantity-limited policy instead of the traditional exhaustive lot-sizing policy. It is important to remark that this gain can be achieved without the need of purchasing additional resources. Furthermore, the quantity-limited policy is easy to implement, augmenting the exhaustive lot-sizing policy by a single "knob" for each product that can tune performance to be efficient (large quantity limits and long production runs) or fair (small quantity limits and short production runs). Finally, we wish to stress the organizational advantages of the quantity-limited lot-sizing policy such as facilitation of maintenance scheduling, workforce planning, purchasing of raw material, scheduling of subsequent processes and shipment of finished products.

4.A Application of Rouché's theorem

The present appendix is based on [P1]. Apart from the results discussed over here, [P1] dilates upon the significance of Theorem 4.A.4 in the analysis of queueing systems and provides some examples of (heavy-tailed) discrete distributions for which the classical approach fails, but to which our result can still be applied. In the vast majority of queueing problems to which Rouché's theorem is applied, the analytic function of interest is given by $z^s - A(z)$, where $s \in \mathbb{N}$ and $A(z)$ is the PGF of a nonnegative discrete random variable

A. Denoting $\mathbb{P}(A = j)$ by a_j, we have that

$$A(z) = \sum_{j=0}^{\infty} a_j z^j, \tag{4.72}$$

which is known to be analytic in the open disk $\{z \in \mathbb{C} : |z| < 1\}$ and continuous up to the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. Note that $A(z)$ is differentiable at $z = 1$ if and only if $\sum_{j=1}^{\infty} j a_{j-1} z^{j-1} < \infty$. If $A(z)$ is differentiable at $z = 1$, it is differentiable at z for all $z \in \mathbb{C}$ with $|z| = 1$.

Let us first state Rouché's theorem (see, e.g., p. 116 of Titchmarsh [209]):

Theorem 4.A.1 (Rouché) *Let the bounded region D have as its boundary a simple closed contour C. Let $f(z)$ and $g(z)$ be analytic both in D and on C. Assume that $|f(z)| < |g(z)|$ on C. Then $f(z) - g(z)$ has in D the same number of zeros as $g(z)$, all zeros counted according to their multiplicity.* □

When $A(z)$ has a radius of convergence larger than one, we can prove the following result concerning the number of zeros on and within the unit circle of $z^s - A(z)$ by using Rouché's theorem:

Lemma 4.A.2 *Let $A(z)$ be a* PGF *that is analytic in $|z| \leq 1+\nu$, for some $\nu > 0$. Assume that $A'(1) < s$, $s \in \mathbb{N}$. Then the function $z^s - A(z)$ has exactly s zeros in $|z| \leq 1$.*

Proof. Define the functions $f(z) := A(z)$, $g(z) := z^s$. Because $f(1) = g(1)$ and $f'(1) = A'(1) < s = g'(1)$, we have, for sufficiently small $\epsilon > 0$,

$$f(1+\epsilon) < g(1+\epsilon). \tag{4.73}$$

Consider all z with $|z| = 1 + \epsilon$. By the triangle inequality and (4.73) we have that

$$|f(z)| \leq \sum_{j=0}^{\infty} a_j |z|^j \;=\; f(1+\epsilon) < g(1+\epsilon) = |g(z)|, \tag{4.74}$$

and hence $|f(z)| < |g(z)|$. Because both $f(z)$ and $g(z)$ are analytic for $|z| \leq 1+\epsilon$, Rouché's theorem tells us that $g(z)$ and $f(z) - g(z)$ have the same number of zeros in $|z| \leq 1 + \epsilon$. Letting ϵ tend to zero yields the proof. □

The application of Lemma 4.A.2 is limited to the class of functions $A(z)$ with a radius of convergence larger than 1. In case $A(z)$ has radius of convergence 1, the results below can be applied. Before we, however, present our main result, we first prove a result on the number and location of zeros of $z^s - A(z)$ on the unit circle. We define the period p of a series $\sum_{-\infty}^{\infty} b_j z^j$ as the largest integer for which $b_j = 0$ whenever j is not divisible by p.

Lemma 4.A.3 *Let $A(z)$ be a* PGF *of some nonnegative discrete random variable with $A(0) > 0$. Assume $A(z)$ is differentiable at $z = 1$ and $A'(1) < s$, where s is a positive integer. If $z^s - A(z)$ has period p, then $z^s - A(z)$ has exactly p zeros on the unit circle given by the p-th roots of unity $\tau_k = \exp(2\pi i k/p)$, $k = 0, 1, \ldots, p-1$. In each of these zeros, the derivative of $z^s - A(z)$ does not vanish.*

Proof. Obviously, any zero ξ of $z^s - A(z)$ with $|\xi| = 1$ is simple, since $|A'(\xi)| \leq A'(|\xi|) = A'(1) < s$ and, thus, $s\xi^{s-1} - A'(\xi) \neq 0$. Furthermore, for any z with $|z| = 1$, $|A(z)| = A(1)$ iff $z^k = 1$ whenever $a_k > 0$. This easily follows from the fact that $|a_0 + a_k z^k| < a_0 + a_k$

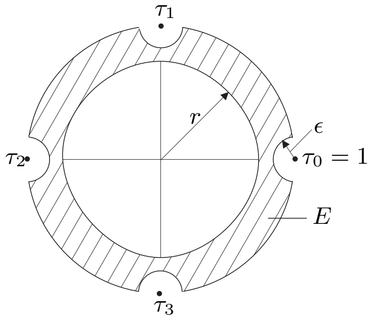


Figure 4.2: Graphical representation of the compact set E.

if $z^k \neq 1$. So, for z with $|z| = 1$ and $A(z) - z^s = 0$ it follows that $z^k = 1$ for all k with $a_k > 0$, and $z^s = 1$. This implies that $z^p = 1$, which completes the proof. □

Note that the requirement $a_0 = A(0) > 0$ involves no essential limitation: If a_0 were zero we would replace the distribution $\{a_i\}_{i\geq 0}$ by $\{a_i^*\}_{i\geq 0}$ where $a_i^* = a_{i+m}$, a_m being the first non-zero entry of $\{a_i\}_{i\geq 0}$, and a corresponding decrease in s according to $s^* = s - m$. In this case, we could thus rewrite $z^s - A(z)$ as $z^m(z^{s-m} - A^*(z))$.

We are now in a position to give the main result:

Theorem 4.A.4 *Let $A(z)$ be a* PGF *of some nonnegative discrete random variable with $A(0) > 0$. Assume $A(z)$ is differentiable at $z = 1$ and $A'(1) < s$, where s is a positive integer. Also, let $z^s - A(z)$ have period p. Then the function $z^s - A(z)$ has p zeros on the unit circle given by $\tau_k = \exp(2\pi ik/p)$, $k = 0, 1, \ldots, p-1$ and exactly $s-p$ zeros in $|z| < 1$.*

Proof. Lemma 4.A.3 tells us that $F(z) = z^s - A(z)$ has p equidistant zeros *on* the unit circle, and so it remains to prove that this function has exactly $s - p$ zeros *within* the unit circle. Thereto, define, for $N \in \mathbb{N}$, the truncated PGF

$$A_N(z) = \sum_{j=0}^{N-1} a_j z^j + \sum_{j=N}^{\infty} a_j z^N, \tag{4.75}$$

where N is a multiple of p. Then $F_N(z) = z^s - A_N(z)$ has obviously s zeros in $z \in D = \{z \in \mathbb{C} : |z| \leq 1\}$, since $A_N(z)$ is a polynomial satisfying $A_N'(1) < s$, and Lemma 4.A.2 thus applies. By Lemma 4.A.3 we know that $F_N(z)$ has p simple and equidistant zeros on the unit circle. We further have that

$$|A(z) - A_N(z)| \;\leq\; 2\sum_{j=N}^{\infty} a_j, \qquad |z| \leq 1, \tag{4.76}$$

$$|A'(z) - A_N'(z)| \;\leq\; 2\sum_{j=N}^{\infty} ja_j, \qquad |z| \leq 1. \tag{4.77}$$

Thus, $A_N(z)$ and $A_N'(z)$ *converge uniformly* to $A(z)$ and $A'(z)$ on $z \in D$, respectively. Moreover, if $G : D \to \mathbb{C}$ is continuous, then $G(A_N(z))$ is uniformly convergent to $G(A(z))$ on $z \in D$.

Take some value z on $C = \{z \in \mathbb{C} : |z| = 1\}$. If for all $n \in \mathbb{N}$ there is a $z_n \in D$ with $0 < |z - z_n| < \frac{1}{n}$ and $F(z_n) = 0$, then $F(z) = 0$ and

$$F'(z) = \lim_{n\to\infty} \frac{F(z_n) - F(z)}{z_n - z} = 0. \tag{4.78}$$

However, this is impossible by Lemma 4.A.3. Hence, there is an $\eta > 0$ such that $F(\xi) \neq 0$ for all $\xi \in D(z, \eta) := \{\xi \in D : 0 < |\xi - z| < \eta\}$. Since C is compact, it can be covered by finitely many $D(z, \eta)$'s. Hence, there is a $0 < r < 1$ such that $F(z)$ has no zeros in $r \leq |z| < 1$.

Now we prove that for large N the function $F_N(z)$, as the function $F(z)$, has no zeros in $r \leq |z| < 1$. Thereto, we show that there is an $\epsilon > 0$ and $M \in \mathbb{N}$ such that $F_N(z) \neq 0$ for all $N \geq M$ and $0 < |z - \tau_k| < \epsilon$, $k = 0, 1, \ldots, p - 1$. Because $F'(z)$ is continuous and $F'_N(z)$ converges uniformly to $F'(z)$ on $z \in D$, there are $\epsilon > 0$ and $M \in \mathbb{N}$ such that for $k = 0, 1, \ldots, p - 1$,

$$|F'_N(z) - F'(\tau_k)| < \delta < |F'(\tau_k)|, \qquad 0 < |z - \tau_k| < \epsilon, \quad N \geq M. \tag{4.79}$$

Furthermore, we have for $k = 0, 1, \ldots, p - 1$,

$$|F_N(z) - F'(\tau_k)(z - \tau_k)| = \Big| \int_{[\tau_k, z]} (F'_N(s) - F'(\tau_k))\mathrm{d}s \Big|, \tag{4.80}$$

where the integration is carried out along the straight line that connects τ_k and z. Hence, for $0 < |z - \tau_k| < \epsilon$ and $N \geq M$, we obtain for $k = 0, 1, \ldots, p - 1$,

$$\Big| \int_{[\tau_k, z]} (F'_N(s) - F'(\tau_k))\mathrm{d}s \Big| \leq |z - \tau_k| \max_{s \in [\tau_k, z]} |F'_N(s) - F'(\tau_k)| < |z - \tau_k|\delta. \tag{4.81}$$

So, it follows that for $0 < |z - \tau_k| < \epsilon$ and $N \geq M$ for $k = 0, 1, \ldots, p - 1$,

$$\begin{aligned} |F_N(z)| &= |F_N(z) - F'(\tau_k)(z - \tau_k) + F'(\tau_k)(z - \tau_k)| \\ &\geq |F'(\tau_k)||z - \tau_k| - |F_N(z) - F'(\tau_k)(z - \tau_k)| \\ &> (|F'(\tau_k)| - \delta)|z - \tau_k| > 0. \end{aligned} \tag{4.82}$$

Since $F_N(z)$ converges uniformly to $F(z)$ and $F(z) \neq 0$ on the compact set (see Figure 4.2)

$$E = \{z \in \mathbb{C} : r \leq |z| \leq 1\} \setminus \bigcup_{k=0}^{p-1} D(\tau_k, \epsilon), \tag{4.83}$$

there exists a $K \in \mathbb{N}$ such that $F_N(z) \neq 0$ for all $N \geq K$ and $z \in \mathbb{C}$ with $r \leq |z| < 1$. Hence, for all $N \geq K$ the number of zeros of $F_N(z)$ with $|z| < r$ is equal to $s - p$. This number can be expressed by the argument principle (see, e.g., Titchmarsh [209]) as follows

$$s - p = \frac{1}{2\pi i} \oint_{|z|=r} \frac{F'_N(z)}{F_N(z)} \mathrm{d}z. \tag{4.84}$$

The integrand converges uniformly to $F'(z)/F(z)$, and thus

$$\frac{1}{2\pi i} \oint_{|z|=r} \frac{F'(z)}{F(z)} \mathrm{d}z = \lim_{N\to\infty} \frac{1}{2\pi i} \oint_{|z|=r} \frac{F'_N(z)}{F_N(z)} \mathrm{d}z = s - p. \tag{4.85}$$

Hence, the number of zeros of $F(z)$ with $|z| < r$ is also $s - p$. This completes the proof. □

Chapter 5

Miscellaneous results

This concluding chapter is divided into two parts. Section 5.1 examines the behavior of the exhaustive and gated discipline for the general class of renewal arrival processes, which is motivated by measurements in practice. Section 5.2 is devoted to an alternative lot-sizing policy, the lowest inventory policy, being the optimal policy in a two-product Markovian system without setup times.

5.1 General arrival processes

Many queueing models dictate that the interarrival times are independent and exponentially distributed; the MVA framework presented in Chapter 2 relies on these assumptions as well. The author has, however, been involved in several industry projects [P21; 185; 211; 215] - mainly from process industry - which show that the (single-item) demand distributions may significantly deviate from the exponential distribution (recall that the interarrival distribution in a queueing model represents the demand distributions in the corresponding production system). Related is the study of Inman [124], who presents data from two automotive body welding lines assessing the validity of exponential and independence assumptions for a wide range of variables. The most important conclusions for this section are that assuming exponential interarrival times is not appropriate, whereas the independence assumptions do appear to be valid. Although the study of [124] and our projects [P21; 185; 211; 215] represent only a minuscule sample of production systems, they do show that the analysis of queueing systems with general (renewal) arrival processes is of practical interest.

Motivated by these observations the present section studies exhaustive and gated polling systems with general arrival processes. So far, hardly any (exact) results have been derived for polling systems with general arrival processes apart from stability conditions and some mean value results for global performance measures such as cycle times (see, e.g., [33] and [81]). The main contribution of the present section, which is based on parts of [P6; P7; P8; P11], is three-fold. First, we explain the use of the algorithm developed in Chapter 4 in case of the exhaustive policy under the assumption of general arrival processes (see Subsection 5.1.2). Second, Subsection 5.1.3 proposes a method to rigorously prove heavy-traffic limits for the expected waiting time in case of increasing load. Third, we provide a conjecture regarding the behavior of systems with renewal arrivals in case of increasing deterministic setup times; support for this conjecture is given by numerical experiments (see Subsection 5.1.4). Below we first give a brief model description.

5.1.1 Model description

The basic N-queue polling system described in Chapter 2 is studied, where we assume that service at each queue is according to the exhaustive or the gated discipline. However, the Poisson arrival processes introduced in Chapter 2 are generalized implying that customers arrive at all queues according to independent general renewal processes. The mean and second moment of the interarrival times are denoted by $\mathbb{E}[A_i]$ and $\mathbb{E}[A_i^2]$, $i = 1, 2, \ldots, N$, respectively. Furthermore, the arrival rate at queue i is denoted by $\lambda_i := 1/E[A_i]$.

5.1.2 General traffic settings

In Chapter 4 we have presented an approximate algorithm for the k-limited policy (which is directly applicable to the exhaustive policy as well) under the most general imaginable assumptions, i.e., general number of queues each with their own service limit in an environment with generally distributed arrival, service and setup times. Recall that within the algorithm approximations for various distributions, such as the interarrival distributions, are obtained by fitting phase-type distributions on the first two moments (see [P8]). It is, however, possible to make more sophisticated use of phase-type distributions by fitting more moments or by approximating the shape of the underlying distribution (see, e.g., [126; 219]). In this respect, it is important to remark that the phase-type distributions used in our fitting approach - the Coxian distributions and the Erlang distributions with the same scale parameters - are *dense* in the class of all distributions on $[0, \infty)$. This means that for any distribution $F(\cdot)$ on $[0, \infty)$ a sequence of phase-type distributions can be constructed which pointwise converge at the points of continuity of $F(\cdot)$ (see [189; 190]).

5.1.3 Increasing load

We now focus on the heavy-traffic behavior of gated polling models, in which the arrival process at each of the queues follows a general renewal process. In case of Poisson arrivals, rigorous proofs for heavy-traffic limits can be obtained for models that possess a multi-type branching process structure (see Chapter 2). In case of renewal arrivals and a general number of queues, heavy-traffic limits have only been obtained on the basis of conjectures [66; 67; 175]. In this subsection, which is based on Van der Mei and Winands [P7], we study a method to derive rigorous proofs for heavy-traffic asymptotics in gated polling models with a general number of queues under the assumption of general renewal arrivals. The approach has its origin in [P6], where we study systems with *zero* setup times. At face value the extension to *nonzero* setup times may seem a small one, however this extension impels us to, considerably, modify and extend the analysis in [P6] as done below.

We start our analysis by extending a result of Bertsimas and Mourtzinou [39] to general setup time distributions, which yields a set of linear equations for the variance of the cycle times for polling models with renewal arrivals in heavy traffic. Exploiting the similarities of this set with the corresponding set for systems with Poisson arrivals yields a closed-form expression for the asymptotic pseudo-conservation law for systems with renewal arrivals in heavy traffic. Subsequently, by taking the proper heavy-traffic limits of this set in combination with the derived pseudo-conservation law we obtain explicit closed-form expressions for the mean asymptotic scaled waiting time in heavy traffic. We present - in the interest of space - detailed proofs only for the gated policy, but we want to stress that the approach is also readily applicable to the exhaustive policy.

Throughout, for each variable x that is a function of ρ, we denote its values *evaluated at* $\rho = 1$ by $\hat{x}$. Furthermore, we use the notation that $h(x) \backsim g(x)$ as $x \uparrow a$ means that $\lim_{\rho \uparrow a} h(x)/g(x) = 1$. Our main interest is in the behavior of the mean waiting time $\mathbb{E}[W_i]$ in heavy traffic, i.e., as ρ tends to 1. It goes without saying that, in heavy traffic, all queues become unstable and, thus, $\mathbb{E}[W_i]$ tends to infinity for all i. To be precise, $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$, for $i = 1, 2, \ldots, N$,

$$\mathbb{E}[W_i] = \frac{\mathbb{E}[W_i^*]}{1-\rho} + o((1-\rho)^{-1}), \qquad \rho \uparrow 1, \tag{5.1}$$

where $g(x) = o(f(x))$ means that $g(x)/f(x) \to 0$ as $x \uparrow 1$. For the validity of the statement that $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$, we refer to [P6; P7]. The main result of the present subsection is the following.

Theorem 5.1.1 *For $i = 1, 2, \ldots, N$,*

$$\mathbb{E}[W_i^*] = \frac{(1+\hat{\rho}_i)}{2}\left(\frac{\sigma^2}{\sum_{j=1}^N \hat{\rho}_j(1+\hat{\rho}_j)} + \mathbb{E}[S]\right), \tag{5.2}$$

with

$$\sigma^2 := \sum_{i=1}^N \hat{\lambda}_i \left(\mathbb{V}ar[B_i] + \hat{\rho}_i^2 \mathbb{V}ar[\hat{A}_i]\right). \tag{5.3}$$

□

Here, the limit is taken such that the arrival rates are increased, while keeping the service time distributions fixed, and keeping the distributions of the interarrival times A_i, $i = 1, \ldots, N$, fixed up to a common scaling constant ρ (i.e., $A_i =_d \rho IA_i$, where IA_i, $i = 1, \ldots, N$, are the interarrival times at $\rho = 1$). Notice that in the case of Poisson arrivals we have $\sigma^2 = \mathbb{E}[B^2]/\mathbb{E}[B]$.

5.1.3.A Proof of Theorem 5.1.1

We start with reviewing and extending the asymptotic results of Bertsimas and Mourtzinou [39] for gated polling systems with arbitrary renewal arrival processes in heavy traffic. Starting point of our analysis is the following expression for the mean waiting time of each customer class as $\rho \uparrow 1$, for $i = 1, 2, \ldots, N$ (cf. [39]),

$$\mathbb{E}[W_i] \backsim \frac{(1+\rho_i)}{2}\left(\frac{\mathbb{V}\mathrm{ar}[C_i]}{\mathbb{E}[C_i]} + \mathbb{E}[C_i]\right) + \frac{(c_{A_i}^2 - 1)\mathbb{E}[B_i]}{2}, \tag{5.4}$$

where $c_{A_i}^2$ is the squared coefficient of variation of the interarrival time for queue i. The mean cycle lengths $\mathbb{E}[C_i]$ can be shown to be independent of the queue involved and are given by, for $i = 1, 2, \ldots, N$ and $\rho < 1$ (see, e.g, [33]),

$$\mathbb{E}[C_i] = \frac{\mathbb{E}[S]}{1-\rho}, \tag{5.5}$$

whereas the variances of the cycle lengths $\mathbb{V}\mathrm{ar}[C_i]$, $i = 1, \ldots, N$, can generally not be obtained in closed form and do depend on the queue involved.

Bertsimas and Mourtzinou [39] prove that the N unknowns $\mathbb{V}\mathrm{ar}[C_i]$, $i = 1, \ldots, N$, satisfy the following set of N linear equations, for $i = 1, 2, \ldots, N$ as $\rho \uparrow 1$,

$$\begin{aligned}
&\left(\frac{1+2\rho_i-\rho_i^3}{2(1+\rho_i)} - \sum_{l=1}^{i-1} F_{i,l}^{(i)} - \sum_{l=i+1}^{N} E_{l,i}^{(i)}\right)\mathbb{V}\mathrm{ar}[C_i] \\
-&\left(\frac{1}{2(1+\rho_i)} + \sum_{l=1}^{i-1} F_{i,l}^{(i+1)} + \sum_{l=i+1}^{N} E_{l,i}^{(i+1)}\right)\mathbb{V}\mathrm{ar}[C_{i+1}] \\
&- \sum_{k\neq i,i+1}\left(\sum_{l=1}^{i-1} F_{i,l}^{(k)} + \sum_{l=i+1}^{N} E_{l,i}^{(k)}\right)\mathbb{V}\mathrm{ar}[C_k] \backsim \frac{H_i\rho_i}{1+\rho_i} + \sum_{l=1}^{i-1} F_{i,l}^{(0)} + \sum_{l=i+1}^{N} E_{l,i}^{(0)},
\end{aligned} \tag{5.6}$$

where the constant H_i is given by, for $i = 1, 2, \ldots, N$, $\rho < 1$,

$$H_i := \lambda_i \mathbb{E}[C_i] \left(\mathbb{V}\text{ar}[B_i] + \rho_i^2 \mathbb{V}\text{ar}[A_i] \right) + \mathbb{V}\text{ar}[S_{i+1}], \tag{5.7}$$

and where the coefficients $E_{i,j}^{(k)}$ and $F_{i,j}^{(k)}$ are, as $\rho \uparrow 1$, recursively defined by,

$$E_{i,j}^{(0)} \backsim (a_i - \rho_i e_j) E_{i-1,j}^{(0)} - a_i f_j E_{i-1,j+1}^{(0)} + f_j E_{i,j+1}^{(0)} + \frac{H_{i-1}\rho_i}{a_{i-1}\rho_{i-1}}, \quad \text{for} \quad i - j = 2, \tag{5.8}$$

and for $k = 1, 2, \ldots, N$ as $\rho \uparrow 1$,

$$E_{i,j}^{(k)} \backsim (a_i - \rho_i e_j) E_{i-1,j}^{(k)} - a_i f_j E_{i-1,j+1}^{(k)} + f_j E_{i,j+1}^{(k)}, \qquad \text{for} \quad i - j = 2, \tag{5.9}$$

and for $k = 0, 1, \ldots, N$ as $\rho \uparrow 1$,

$$E_{i,j}^{(k)} \backsim (a_i - \rho_i e_j) E_{i-1,j}^{(k)} - a_i f_j E_{i-1,j+1}^{(k)} + f_j E_{i,j+1}^{(k)}, \qquad \text{for} \quad i - j \geq 3, \tag{5.10}$$

$$F_{i,j}^{(k)} \backsim (a_i - \rho_i e_j) F_{i-1,j}^{(k)} - a_i f_j F_{i-1,j+1}^{(k)} + f_j F_{i,j+1}^{(k)}, \qquad \text{for} \quad i - j \geq 2, \tag{5.11}$$

with initial conditions, for $j = 1, 2, \ldots, N$ as $\rho \uparrow 1$,

$$E_{j,j}^{(0)} \backsim H_j, \tag{5.12}$$

$$E_{j,j}^{(k)} \backsim \begin{cases} \rho_j^2, & k = j, \\ 0, & \text{else}, \end{cases} \tag{5.13}$$

and, for $j = 1, 2, \ldots, N - 1$ as $\rho \uparrow 1$,

$$E_{j+1,j}^{(0)} \backsim \frac{H_j \rho_{j+1}}{1 + \rho_j}, \tag{5.14}$$

$$E_{j+1,j}^{(k)} \backsim \begin{cases} \frac{\rho_j (1 + 2\rho_j) \rho_{j+1}}{2(1+\rho_j)}, & k = j, \\ +\frac{\rho_j \rho_{j+1}}{2(1+\rho_j)}, & k = j + 1, \\ 0, & \text{else}. \end{cases} \tag{5.15}$$

Moreover, for $j = 1, 2, \ldots, N$ as $\rho \uparrow 1$,

$$F_{j,j}^{(0)} \backsim \frac{H_j \rho_j}{1 + \rho_j}, \tag{5.16}$$

$$F_{j,j}^{(k)} \backsim \begin{cases} \frac{\rho_j (1 + 2\rho_j + 2\rho_j^3)}{2(1+\rho_j)}, & k = j, \\ -\frac{\rho_j}{2(1+\rho_j)}, & k = j + 1, \\ 0, & \text{else}. \end{cases} \tag{5.17}$$

and, for $j = 1, 2, \ldots, N - 1$ as $\rho \uparrow 1$,

$$F_{j+1,j}^{(0)} \backsim \frac{e_j \rho_{j+1}}{1 + \rho_j} H_j + \frac{f_j \rho_{j+1}}{1 + \rho_{j+1}} H_{j+1}, \tag{5.18}$$

$$F_{j+1,j}^{(k)} \backsim \begin{cases} \frac{e_j \rho_j (1 + 2\rho_j) \rho_{j+1}}{2(1+\rho_j)}, & k = j, \\ +\frac{e_j \rho_j \rho_{j+1}}{2(1+\rho_j)} + \frac{f_j \rho_{j+1} (1 + 2\rho_{j+1} + 2\rho_{j+1}^3)}{2(1+\rho_{j+1})}, & k = j + 1, \\ -\frac{f_j \rho_{j+1}}{2(1+\rho_{j+1})}, & k = j + 2, \\ 0, & \text{else}, \end{cases} \tag{5.19}$$

where all indices in the above expressions should be read cyclically. Finally, the constants a_i, e_i and f_i are defined as, respectively, for $i = 1, 2, \ldots, N$ as $\rho \uparrow 1$,

$$a_i \backsim \frac{\rho_i(1+\rho_{i-1})}{\rho_{i-1}}, \qquad e_i \backsim \frac{\rho_i}{1+\rho_i} \qquad \text{and} \quad f_i \backsim \frac{1}{a_{i+1}}. \tag{5.20}$$

The complexity of the set (5.6) prevents us from solving it explicitly in general, but we do obtain closed-form expressions in the following cases. First, if we restrict our attention to a specific weighted sum of the solutions for $\mathbb{V}\text{ar}[C_i]$, $i = 1, \ldots, N$, we obtain an explicit closed-form expression immediately leading to the pseudo-conservation law of the model under consideration. Second, we can apply asymptotic expansions to find asymptotically exact closed-form expressions for the dominating factors of $\mathbb{V}\text{ar}[C_i]$ by analyzing a scaled version of (5.6) in combination with the just derived pseudo-conservation law. First, we present two remarks.

Remark 5.1.2 The asymptotic approach expounded in the present section is exact for Poisson processes under any traffic intensity $\rho < 1$, cf. [188] (implying that all $\backsim$-signs could be replaced by =-signs), where we note that in this case, for $i = 1, 2, \ldots, N$,

$$c^2_{A_i} = 1, \tag{5.21}$$

and where the constant H_i reduces to, for $i = 1, 2, \ldots, N$,

$$H_i = \lambda_i \mathbb{E}[C_i]\mathbb{E}[B_i^2] + \mathbb{V}\text{ar}[S_{i+1}]. \tag{5.22}$$

□

Remark 5.1.3 In [39], the identity (5.4) and the set (5.6) are actually derived only for the special case of *deterministic* setup times; details of the derivation in case of *stochastic* setup times leading to (5.4) and (5.6) in full generality are available from the author upon request. □

Asymptotic pseudo-conservation law. By working out an expression for the weighted sum of the solutions for $\mathbb{V}\text{ar}[C_i]$ of the set (5.6), the present subsection derives a pseudo-conservation law for the mean waiting times for the model described in Section 5.1.1 as shown in the following lemma.

Lemma 5.1.4 *As $\rho \uparrow 1$, we have*

$$\sum_{i=1}^{N} \rho_i \mathbb{E}[W_i] \backsim \frac{\rho}{2\mathbb{E}[S]} \sum_{i=1}^{N} H_i + \frac{\mathbb{E}[S]}{2(1-\rho)} \sum_{i=1}^{N} \rho_i(1+\rho_i) + \sum_{i=1}^{N} \rho_i \frac{(c^2_{A_i} - 1)\mathbb{E}[B_i]}{2}. \tag{5.23}$$

Proof. Starting point of our proof is the set of equations for $\mathbb{V}\text{ar}[C_i]$ given by (5.6) in the special case of Poisson arrivals. Recall that in this Poisson case (5.6) is exact under any traffic intensity $\rho < 1$ and H_i is given by (5.22). Notice that the coefficient matrix in the lefthand side of (5.6) is *independent* of H_i and that the righthand side of (5.6) is a *linear* function of H_i implying that the solutions for $\mathbb{V}\text{ar}[C_i]$ of (5.6) are linear functions of H_i as well. That is, for $i = 1, 2, \ldots, N$ and $\rho < 1$,

$$\mathbb{V}\text{ar}[C_i] = g_i\left(H_1, H_2, \ldots, H_N\right), \tag{5.24}$$

where $g_i : \mathbb{R}^N \to \mathbb{R}$ are (unknown) linear functions of $H_1, \ldots, H_N$, i.e., there exist constants $c_{i,j}$, $i, j = 1, \ldots, N$, such that for $i = 1, \ldots, N$,

$$g_i(H_1, \ldots, H_N) = c_{i,0} + \sum_{j=1}^{N} c_{i,j} H_j. \tag{5.25}$$

In order to find a closed-form expression for a weighted sum of these functions, we use the pseudo-conservation law for gated polling systems with Poisson arrivals (see Chapter 2). For $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i \mathbb{E}[W_i] = \frac{\rho}{2\mathbb{E}[S]} \sum_{i=1}^{N} H_i + \frac{\mathbb{E}[S]}{2(1-\rho)} \sum_{i=1}^{N} \rho_i(1+\rho_i). \tag{5.26}$$

For this Poisson case, by using simple balance arguments the mean waiting time at Q_i can be expressed in terms of the first two moments of C_i as follows. For $i = 1, 2, \ldots, N$ and $\rho < 1$,

$$\mathbb{E}[W_i] = \frac{1+\rho_i}{2} \left(\frac{\mathbb{V}\text{ar}[C_i]}{\mathbb{E}[C_i]} + \mathbb{E}[C_i] \right) = \frac{1+\rho_i}{2(1-\rho)} \left(\frac{g_i\left(H_1, H_2, \ldots, H_N\right)(1-\rho)^2}{\mathbb{E}[S]} + \mathbb{E}[S] \right), \tag{5.27}$$

where the last equality follows from application of (5.5) and (5.25). Subsequently, substituting (5.27) into (5.26) yields the following weighted sum of $g_i\left(H_1, H_2, \ldots, H_N\right)$ in the Poisson case. For $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i(1+\rho_i) g_i\left(H_1, H_2, \ldots, H_N\right) = \frac{\rho}{1-\rho} \sum_{i=1}^{N} H_i. \tag{5.28}$$

Returning to the general case of renewal arrivals, (5.6) states that asymptotically $\mathbb{V}\text{ar}[C_i]$, $i = 1, 2, \ldots, N$, satisfy the *same set of linear equations* as in the Poisson case, where the variables H_i, $i = 1, 2, \ldots, N$, are defined as in (5.7). Due to the fact that the coefficient matrix in the lefthand side of (5.6) is a linear invertible mapping in conjunction with the fact that the H_i, $i = 1, 2, \ldots, N$, defined in (5.7), only show up at the right-hand side of (5.6), we have that as $\rho \uparrow 1$, $i = 1 \ldots, N$,

$$\mathbb{V}\text{ar}[C_i] \backsim g_i(H_1, \ldots, H_N) = c_{i,0} + \sum_{j=1}^{N} c_{i,j} H_j, \tag{5.29}$$

where the last equality follows from (5.25). Note that the variables H_i, $i = 1, 2, \ldots, N$, are generally not the same as in the Poisson case. Here, the crucial observation is that the coefficients $c_{i,j}$ in (5.29) are the same as those in the Poisson case (5.25). This immediately implies that (5.28) remains asymptotically true for renewal arrivals, i.e., as $\rho \uparrow 1$,

$$\sum_{i=1}^{N} \rho_i(1+\rho_i) g_i\left(H_1, H_2, \ldots, H_N\right) \backsim \frac{\rho}{1-\rho} \sum_{i=1}^{N} H_i. \tag{5.30}$$

Finally, calling upon (5.4) in combination with (5.29) completes the proof. □

The above pseudo-conservation law is exact for Poisson arrival processes under any traffic intensity $\rho < 1$ and, therewith, generalizes the pseudo-conservation law in gated polling systems with Poisson arrivals [47].

Mean asymptotic scaled waiting times. As mentioned earlier, the set (5.6) can in general not be solved in closed form, but the present subsection finds explicit expressions for the dominating terms of $\mathbb{V}\mathrm{ar}[C_i]$ in heavy traffic. Thereto, we multiply both sides of (5.6) by $(1-\rho)^2$ and let $\rho \uparrow 1$, which renders the corresponding scaled set, for $i = 1, 2, \ldots, N$,

$$\begin{aligned}\left(\frac{1+2\hat{\rho}_i-\hat{\rho}_i^3}{2(1+\hat{\rho}_i)} - \sum_{l=1}^{i-1} F_{i,l}^{(i)} - \sum_{l=i+1}^{N} E_{l,i}^{(i)}\right) \mathbb{V}ar[C_i^*] \\ -\left(\frac{1}{2(1+\hat{\rho}_i)} + \sum_{l=1}^{i-1} F_{i,l}^{(i+1)} + \sum_{l=i+1}^{N} E_{l,i}^{(i+1)}\right) \mathbb{V}ar[C_{i+1}^*] \\ - \sum_{k\neq i,i+1} \left(\sum_{l=1}^{i-1} F_{i,l}^{(k)} + \sum_{l=i+1}^{N} E_{l,i}^{(k)}\right) \mathbb{V}ar[C_k^*] \;=\; 0,\end{aligned} \tag{5.31}$$

where $\mathbb{V}ar[C_i^*]$ represents the variance of the asymptotic scaled i-cycle, i.e., for $i = 1, 2, \ldots, N$,

$$\mathbb{V}ar[C_i^*] = \lim_{\rho\uparrow 1} (1-\rho)^2 \mathbb{V}\mathrm{ar}[C_i], \tag{5.32}$$

where the existence of the limit is guaranteed by the fact that $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$ in conjunction with (5.4) and (5.5). The set (5.31) can be solved up to some unknown scaling factor $c \in \mathbb{R}$ as shown in the following lemma.

Lemma 5.1.5 *The solution of the set (5.31) is given by, for $i = 1, 2, \ldots, N$,*

$$\mathbb{V}ar[C_i^*] = c, \tag{5.33}$$

with $c \in \mathbb{R}$.

Proof. One can verify that in (5.31), for $i = 1, 2, \ldots, N$,

$$\frac{2\hat{\rho}_i-\hat{\rho}_i^3}{2(1+\hat{\rho}_i)} - \sum_{k=1}^{N}\left(\sum_{l=1}^{i-1} F_{i,l}^{(k)} + \sum_{l=i+1}^{N} E_{l,i}^{(k)}\right) = 0, \tag{5.34}$$

which shows that (5.33) is indeed a solution of the *homogeneous* set (5.31). Either by elementary, but tedious, row and column operations or by quoting from [P6] we observe that the rank of the coefficient matrix of (5.31) equals $N-1$, which completes the proof. □

Since the dimension of the null space of the coefficient matrix of (5.31) equals one, adding a single *non-homogeneous* equation would render a unique solution for the unknown scaling factor c. This additional equation can be readily obtained from a scaled version of the pseudo-conservation law (5.23) as done in the lemma below.

Lemma 5.1.6 *For $i = 1, 2, \ldots, N$, $\mathbb{V}ar[C_i^*]$ is given by*

$$\mathbb{V}ar[C_i^*] = \frac{\mathbb{E}[S]\sigma^2}{\sum_{i=1}^{N}\hat{\rho}_i(1+\hat{\rho}_i)}. \tag{5.35}$$

Proof. Via Lemma 5.1.5 in combination with (5.4) and (5.5), one obtains the mean asymptotic scaled waiting times, for $i = 1, 2, \ldots, N$,

$$\mathbb{E}[W_i^*] = \frac{(1+\hat{\rho}_i)}{2}\left(\frac{c}{\mathbb{E}[S]} + \mathbb{E}[S]\right), \tag{5.36}$$

which satisfies a scaled version of the pseudo-conservation law (5.23). That is, multiplying both sides of (5.23) by $(1-\rho)$ and letting $\rho \uparrow 1$ yields

$$\sum_{i=1}^{N} \hat{\rho}_i \mathbb{E}[W_i^*] = \frac{\sigma^2}{2} + \frac{\mathbb{E}[S]}{2} \sum_{i=1}^{N} \hat{\rho}_i (1+\hat{\rho}_i), \tag{5.37}$$

where we have used the definition of σ^2 as given in (5.3). Combining (5.36) and (5.37) completes the proof. □

Lemma 5.1.6 has the following immediate consequence for the mean asymptotic scaled waiting time $\mathbb{E}[W_i^*]$ at each of the queues, which is the main result of the present subsection.

Corollary 5.1.7 *For* $i = 1, 2, \ldots, N$,

$$\mathbb{E}[W_i^*] = \frac{(1+\hat{\rho}_i)}{2} \left(\frac{\sigma^2}{\sum_{j=1}^{N} \hat{\rho}_j (1+\hat{\rho}_j)} + \mathbb{E}[S] \right). \tag{5.38}$$

□

For Poisson arrival processes, the result in Corollary 5.1.7 has been obtained before in the literature, see, e.g., [162]. For general renewal arrivals, only conjectures [66; 67; 175] have been known so far and as such our approach is the first to give a rigorous proof of these conjectures. In [P6] simple closed-form approximations for the mean waiting times in stable systems based on (5.38) are suggested and tested, which show that such approximations are accurate when the load is 90% or more. Furthermore, [P6] argues how the results could be used to prove that the correlations between successive station times in gated systems with renewal arrivals converge to one as the load tends to one (which we have proved before in Chapter 3 for the special case of Poisson arrivals). The method might be extended to derive heavy-traffic results for the complete waiting-time distributions as well. That is, decomposition results for the waiting time distributions obtained in [40] may form a starting point to obtain such results, opening up a very challenging area for further research.

5.1.4 Increasing setup times

In the present subsection we probe exhaustive polling systems with general renewal arrival processes when the deterministic setup times tend to infinity. In the exact MVA analysis of Chapter 3, we have assumed that the arrival processes follow Poisson distributions. If we take a second look at the intuitive interpretation of these results, one would however expect that also in case of general (renewal) arrival processes the polling system converges to a deterministic cyclic system when the setup times tend to infinity. Unfortunately, the techniques used throughout in Chapter 3 rely heavily on the Poisson assumption, i.e., we have exploited MVA results for polling systems with *finite* setup times obtained under the Poisson assumption and, subsequently, we have shown that significant simplifications result as the setup times tend to *infinity*. However, corresponding polling results for general arrival processes are not known.

To numerically test the above conjecture for general arrival processes, we have performed a couple of simulation experiments of exhaustive polling systems. We consider a symmetric polling system with 3 queues, where the service times are exponential with mean 0.25. Interarrival times have mean 1 and the corresponding squared coefficient of variation $c_{A_i}^2$ is varied between 0.25, 0.5, 1 and 2. In order to obtain a distribution for these interarrival times, we fit a phase-type distribution on the first two moments as described in [P11]. In

General arrival processes				
	$c^2_{A_i} = 0.25$	$c^2_{A_i} = 0.5$	$c^2_{A_i} = 1$	$c^2_{A_i} = 2$
$S_i = 1$	0.121	0.167	0.259	0.444
$S_i = 10$	0.012	0.017	0.026	0.044
$S_i = 50$	0.002	0.003	0.005	0.009
$S_i = 100$	0.001	0.002	0.003	0.004

Table 5.1: Coefficient of variation of the (scaled) number of customers at queue i at a polling instant of queue i.

case the squared coefficient of variation equals 1 the arrival process is approximated by a Poisson process and this case is included as benchmark.

Table 5.1 shows the coefficient of variation of the (scaled) number of customers X_i at queue i at a polling instant of queue i for varying values of the marginal deterministic setup times S_i in a cycle. From this table, we clearly see that the coefficient of variation approaches zero when the setup times tend to infinity. It goes without saying that a highly variable arrival process has a negative impact on how "fast" the limiting behavior is approached. Via Chebyshev's inequality (see, e.g., [183]) we know that a random variable with zero variance follows a deterministic distribution and, therefore, this observation provides empirical evidence for the fact that the scaled number of customers at queue i at a polling instant of queue i becomes deterministic. Therefore, it confirms the validity of our conjecture that the polling system converges to a deterministic cyclic system as the setup times increase to infinity. Obviously, a more extensive test bed is needed to test our hypothesis more rigorously, but without doubt extending our work to general arrival processes is a very interesting topic for further research.

Concluding we can say that the asymptotic analyses of the present section confirm the qualitative observations made within the mathematically rigorously developed MVA framework of Chapter 3 for the cases of high utilizations due to either high load or large setup times. Although the actual numbers computed within the MVA framework may deviate from practical measurements (due to the Poisson assumption), the predicted behavior of the system is close to reality.

5.2 Lowest inventory policy

As seen in Chapter 2, for general instances of the SELSP there are no structural results known to be satisfied by an optimal policy. However, for the following occurrence of the SELSP the optimal policy is known. Consider a system without setup times and costs and with two completely *identical* products with Poisson demand rate λ, exponential production times with rate μ and identical relevant cost factors across the products. The class of base-stock policies is optimal for this system (see Zheng and Zipkin [223]); a production order is placed when the inventory position (physical inventory plus the stock on order minus backorders) of a product falls below a pre-defined target inventory level. Due to the assumption of identical products, these target inventory levels are identical for both products. Subsequently, these production orders queue up at the machine, where we optimize the decision which product to produce next. The optimal decision turns out to be global and should be made according to the *lowest inventory* policy: the machine produces the product having the lowest net inventory position (physical inventory minus backorders). Furthermore, ties are resolved randomly and this policy is applied preemptively. The optimality result only depends on the symmetry of the products and the convexity of the cost function (see Menich and Serfozo [168]).

Although we disregard setup times, our modeling approach allows for the inclusion of setup costs representing external setups - which can be executed off-line - in opposition to setup times which fully occupy the machine. Furthermore, we want to stress that the implementation and control of the lowest inventory policy requires *global* information, which is a serious drawback in production environments. Moreover, the extension of the lowest inventory policy to general asymmetric systems or general systems with setup times is certainly not trivial (see, e.g., [119]). That is, there exists no comprehensive account of the optimal policy in asymmetric systems and systems with setup times; however, index policies (see, e.g., [171]) and threshold policies (see, e.g., [48]), respectively, seem to work well in such extensions.

It can be easily shown that the shortfall distributions for this lowest inventory policy do not depend on the base-stock levels and equal the queue length distributions in a standard queueing model (cf. the fixed-sequence base-stock policies introduced in Chapter 1). In this queueing model, there are two Poisson arrival streams with rate λ and a single exponential server with rate μ, which always works on the longer queue, i.e., on the product having the most outstanding production orders. Therefore, the lowest inventory policy is also known as the *longest queue* policy. In the remainder of the present section, we adopt the nomenclature as used in the field of queueing theory, e.g., we talk about queue length distribution instead of shortfall distribution. For other studies on this longest queue policy, we refer to [26; 69; 97; 128; 223; 225] and the references therein.

The main contribution of the present section, which is an excerpt of [P2; P3], is as follows. The queue length distribution in case of the longest queue policy can be computed by modeling the system as a quasi-birth-and-death (QBD) process and using matrix-geometric techniques. This approach requires the determination of a so-called rate matrix $\mathbf{R}$, which in most applications needs to be computed by some iterative algorithm. In Subsection 5.2.1 we present, however, a general class of QBD processes for which $\mathbf{R}$ can be determined explicitly, based on probabilistic arguments. The longest queue policy falls within this class. For models falling within this class, we reformulate the probabilistic interpretations of the fundamental matrices in terms of Bernoulli excursions, leading to explicit expressions for the matrix elements in terms of hypergeometric functions.

5.2.1 QBD processes with an explicit rate matrix

Consider a continuous-time Markov process $\{X(t) : t \in \mathbb{R}^+\}$ on the two-dimensional state space $\mathbb{Z}^+ \times \{0, \ldots, d\}$, where $\mathbb{R}^+$ ($\mathbb{Z}^+$) denotes the set of nonnegative reals (integers) and d may be finite or infinite. The set $\{(i,j) : j \in \{0, \ldots, d\}\}$ is called *level* i, $i \in \mathbb{Z}^+$, whereas the second dimensional component is called the *phase*. Define $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$, $\boldsymbol{\pi}_i = (\pi(i,0), \pi(i,1), \ldots, \pi(i,d))$ and $\pi(i,j) = \lim_{t\to\infty} \mathbb{P}[X(t) = (i,j)]$, $i \in \mathbb{Z}^+$, $j \in \{0, \ldots, d\}$. The limiting probability vector $\boldsymbol{\pi}$ is the stationary distribution for the stochastic process $\{X(t) : t \in \mathbb{R}^+\}$. We shall assume throughout that this process is irreducible and positive recurrent, so that the invariant probability vector is uniquely determined by solving $\boldsymbol{\pi}\mathbf{Q} = 0$ and $\boldsymbol{\pi}\mathbf{e} = 1$, where $\mathbf{Q}$ is the infinitesimal generator matrix for the process and $\mathbf{e}$ is a column vector of appropriate dimension containing all ones.

A Markov process is called a homogeneous QBD process when one-step transitions are restricted to states in the same level or in two adjacent levels, and the transition rates are assumed to be level-independent. The generator $\mathbf{Q}$ then has the block-tridiagonal structure,

$$\mathbf{Q} = \begin{pmatrix} \mathbf{B}_{00} & \mathbf{B}_{01} & & & \\ \mathbf{B}_{10} & \mathbf{A}_1 & \mathbf{A}_0 & & \\ & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{5.39}$$

The stationary probability vector $\boldsymbol{\pi}$ of the QBD process $\{X(t) : t \in \mathbb{R}^+\}$ with generator $\mathbf{Q}$ satisfies the following standard result (see Latouche and Ramaswami [141]).

Theorem 5.2.1 *Consider the QBD process $\{X(t) : t \in \mathbb{R}^+\}$ with infinitesimal generator $\mathbf{Q}$ in the form of (5.39). Suppose that this stochastic process is irreducible and positive recurrent. Then the stationary distribution $\boldsymbol{\pi}$ is given by*

$$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n \mathbf{R}, \quad n \geq 1, \tag{5.40}$$

where $\mathbf{R}$ is the minimal nonnegative solution of the nonlinear matrix equation,

$$\mathbf{A}_0 + \mathbf{R}\mathbf{A}_1 + \mathbf{R}^2\mathbf{A}_2 = 0. \tag{5.41}$$

□

Determining $\mathbf{R}$ is essentially tantamount to determining the related matrix $\mathbf{G}$ that satisfies (see [141]),

$$\mathbf{A}_0\mathbf{G}^2 + \mathbf{A}_1\mathbf{G} + \mathbf{A}_2 = 0. \tag{5.42}$$

The probabilistic interpretation of the element $r_{j,k}$ of the matrix $\mathbf{R} := [r_{j,k}]_{j,k\in\{0,\dots,d\}}$ is the expected amount of time spent in state $(i+1,k)$ before the first return to any state of level i, expressed in units of the mean sojourn time for the state (i,j), given that the process started in state (i,j), $i \in \mathbb{Z}_{\geq 1}$ (see [141; 170]).

The related matrix $\mathbf{G} := [g_{j,k}]_{j,k\in\{0,\dots,d\}}$ is defined such that

$$g_{j,k} = \mathbb{P}[\tau < \infty, X(\tau) = (i,k) | X(0) = (i+1,j)], \tag{5.43}$$

where τ is the first passage time from the level $i+1$ to the level i, $i \in \mathbb{Z}_{\geq 1}$ (see again [141; 170]). Providing $\{X(t) : t \in \mathbb{R}^+\}$ is irreducible and positive recurrent, it follows that $\tau < \infty$ with probability 1 and, thus,

$$g_{j,k} \;=\; \mathbb{P}[X(\tau) = (i,k) | X(0) = (i+1,j), \tau < \infty]. \tag{5.44}$$

In order to obtain the stationary distribution, one should thus determine the rate matrix $\mathbf{R}$. Several iterative procedures exist for solving (5.41); an overview of such algorithms is given in [140]. For the class of QBD processes described below, we exploit the probabilistic interpretations of the elements $r_{j,k}$ and $g_{j,k}$ to obtain explicit solutions for $\mathbf{R}$.

Special class. Denote by $\langle e_1, e_2\rangle$ a one-step transition of the QBD process from state (i,j) to state $(i+e_1, j+e_2)$. For the special class, possible steps in state (i,j), $i \in \mathbb{Z}_{\geq 1}$, $j \in \{0,\dots,d-1\}$ are horizontal steps $\langle -1, 0\rangle$ and $\langle 1, 0\rangle$, diagonal steps $\langle -1, 1\rangle$ and $\langle 1, 1\rangle$, or a vertical step $\langle 0, 1\rangle$. The exponential rate at which a step occurs is denoted by $r\langle e_1, e_2\rangle$ and the probability that a step occurs is given by

$$\varphi\langle e_1, e_2\rangle := \frac{r\langle e_1, e_2\rangle}{\sum_{\text{steps}} r\langle e_1, e_2\rangle}. \tag{5.45}$$

For presentation reasons, the focus of the present subsection is on the case where d is infinite and where the diagonal elements of A_1 are all equal. Both restrictions are not prohibitive (see [P2; P3]). For the introduced class, $\mathbf{R}$ and $\mathbf{G}$ are of uppertriangular form, i.e., with $r_h \equiv r_{j,j+h}$ and $g_h \equiv g_{j,j+h}$,

$$\mathbf{R} = \begin{pmatrix} r_0 & r_1 & r_2 & \cdots \\ & r_0 & r_1 & \cdots \\ & & r_0 & \cdots \\ & & & \ddots \end{pmatrix}, \qquad \mathbf{G} = \begin{pmatrix} g_0 & g_1 & g_2 & \cdots \\ & g_0 & g_1 & \cdots \\ & & g_0 & \cdots \\ & & & \ddots \end{pmatrix}. \tag{5.46}$$

We now consider the paths involved in the probabilistic interpretations of the elements $r_{j,k}$ and $g_{j,k}$ and decompose the corresponding passage probabilities to derive expressions for these fundamental matrix elements in terms of Bernoulli excursions. Although such expressions can be directly obtained for either the **R** or **G** matrices, it can be helpful (with respect to both the analysis and the presentation) to first consider the derivation of the elements of the **G** matrix and, then, turn to consider the elements of the **R** matrix. The presentation is organized accordingly.

Matrix G. We derive expressions for each element of **G** in (5.46), using the interpretation (5.44). Assume that a time τ the process undergoes its ν-th transition. Our main idea is the decoupling of paths of the QBD process into horizontal and vertical directions.

(i) Consider a path from state $(i+1,j)$ to $(i,j+h)$, $h \geq 0$, that consists of ν steps and that goes from level $i+1$ to i only at the last (ν-th) step. Assume this path contains s $\langle -1,1\rangle$ steps, u $\langle 1,1\rangle$ steps and hence $t = h - s - u$ $\langle 0,1\rangle$ steps.

(ii) We would first like to consider the path in the horizontal direction only. The diagonal steps $\langle -1,1\rangle$ and $\langle 1,1\rangle$ influence both the horizontal and vertical directions. Therefore, we decompose the diagonal steps into

$$\langle -1,1\rangle \;=\; \langle 0,1\rangle \;+\; \langle -1,0\rangle, \tag{5.47}$$

$$\langle 1,1\rangle \;=\; \langle 0,1\rangle \;+\; \langle 1,0\rangle. \tag{5.48}$$

(iii) The decomposition of the diagonal steps ensures that the path contains at least s $\langle -1,0\rangle$ steps and u $\langle 1,0\rangle$ steps. Now denote the total number of $\langle 1,0\rangle$ steps by m. We then know that the total number of $\langle -1,0\rangle$ steps is $m+1$ (including the ν-th step). Furthermore, it should hold that $m \geq \max(u, s-1)$.

(iv) The path then consists of a total number of $\nu = 2m + t + 1$ steps, $2m+1$ of which are in horizontal direction. When we omit the ν-th step, the $2m$ horizontal steps form a Bernoulli excursion (see Takács [202]). The excursion starts at $(i+1,j)$ and consists of m $\langle 1,0\rangle$ steps and m $\langle -1,0\rangle$ steps. Any sequence of steps may occur, as long as there are, at each point during the excursion, at least as many $\langle 1,0\rangle$ steps as $\langle -1,0\rangle$ steps, for otherwise, level i is visited, and the condition in (i) is violated. The number of possible Bernoulli excursions is given by the m-th Catalan number

$$\frac{1}{m+1}\binom{2m}{m}. \tag{5.49}$$

To elucidate the exposition of our derivation, let us illustrate the above procedure with an example. In Figure 5.1 we see a path from state $(i+1,j)$ to $(i,j+5)$ that consists of $\nu = 21$ steps and that goes from level $i+1$ to i only at the ν-th step. We have indicated in Figure 5.1 for each state the number of times it is visited. The path contains two $\langle -1,1\rangle$ steps, one $\langle 1,1\rangle$ step and two $\langle 0,1\rangle$ steps. The diagonal steps are decomposed, leading to the path in Figure 5.2. We then remove the vertical steps to obtain the Bernoulli excursion in Figure 5.3.

The excursion in Figure 5.3 is just one out of the $\frac{1}{10}\binom{18}{9} = 4862$ possible Bernoulli excursions that consist of $2m = 18$ steps. In the procedure we consider just one path from state (i,j) to state $(i,j+h)$, and reduce it to a Bernoulli excursion. We now change perspectives and start from a Bernoulli excursion of length $2m$ and ask ourselves how many paths from (i,j) to state $(i,j+h)$ can be constructed with s $\langle -1,1\rangle$ steps, u $\langle 1,1\rangle$ steps and thus $t = h - s - u$ $\langle 0,1\rangle$ steps. We first need some definitions.

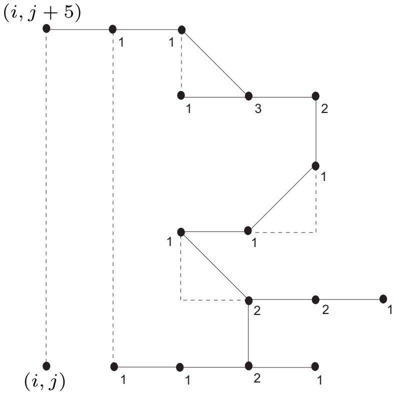


Figure 5.1: Path from state $(i+1, j)$ to $(i, j+5)$.

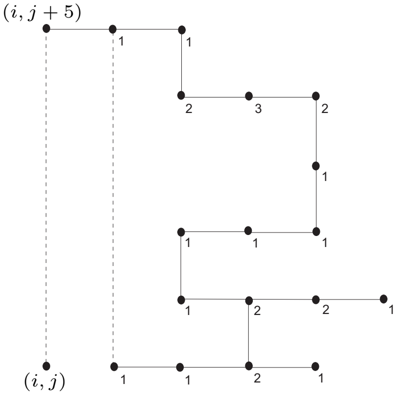


Figure 5.2: Path without diagonal steps.

Definition 5.2.2 *Define $L_h(s, u, m)$ as the number of paths from $(i+1, j)$ to $(i, j+h)$, $h \geq 0$, that consist of $\nu = 2m+1+t$ steps and that go from level $i+1$ to i only at the ν-th step. Assume each path contains s $\langle -1, 1\rangle$ steps, u $\langle 1, 1\rangle$ steps and hence $t = h - s - u$ $\langle 0, 1\rangle$ steps. Let $P_h(s, u, m)$ denote the probability of each such path.* □

Obviously,

$$P_h(s, u, m) = \varphi\langle -1, 1\rangle^s \varphi\langle 0, 1\rangle^t \varphi\langle 1, 1\rangle^u \varphi\langle 1, 0\rangle^{m-u} \varphi\langle -1, 0\rangle^{m+1-s}. \tag{5.50}$$

For $L_h(s, u, m)$ we have the following result.

Lemma 5.2.3 *For values of s and u such that $s+u+t = h$ we have that*

$$L_h(s, u, m) = \frac{1}{m+1}\binom{2m}{m}\binom{m+1}{s}\binom{m}{u}\binom{2m+t}{t}. \tag{5.51}$$

Proof.

(v) Consider a Bernoulli excursion of length $2m$. We will extend the Bernoulli excursion, that describes the path in horizontal direction, by the vertical steps to reconstruct a path from state $(i+1, j)$ to $(i, j+h)$. The vertical steps consist of $s+t+u$ $\langle 0, 1\rangle$ steps. However, because of the decomposition in (ii), s steps should be matched to $\langle -1, 0\rangle$ steps and u steps should be matched to $\langle 1, 0\rangle$ steps. The number of ways to do this is

$$\binom{m+1}{s}\binom{m}{u}. \tag{5.52}$$

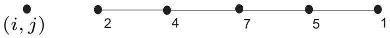


Figure 5.3: Bernoulli excursion.

(vi) This leaves the t original $\langle 0,1\rangle$ steps to be placed. These steps can be placed in every state visited by the Bernoulli excursion. This means that we have $2m+1$ possible states in which the t $\langle 0,1\rangle$ steps should be placed. Note that possibly multiple steps can be placed in one state. Therefore, placing the t $\langle 0,1\rangle$ steps is equivalent to distributing t balls over $2m+1$ bins, and the number of ways to do this is

$$\binom{2m+t}{t}. \tag{5.53}$$

Combining (5.49), (5.52) and (5.53) completes the proof. □

Using (5.44) and $g_h = g_{j,j+h}$ we then arrive at our main result.

Theorem 5.2.4 *For all $h = 0, 1, \ldots$ we have that*

$$g_h = \sum_{s=0}^{h} \sum_{u=0}^{h-s} \sum_{m=\max(u,s-1)}^{\infty} L_h(s,u,m) \cdot P_h(s,u,m), \tag{5.54}$$

with $P_h(s,u,m)$ as in (5.50) *and $L_h(s,u,m)$ as in* (5.51). □

Let us return to the example path in Figures 5.1-5.3. Starting from the Bernoulli excursion in Figure 5.3, the path in Figure 5.1 is just one of the $\binom{10}{2}\binom{9}{1}\binom{20}{2} = 76950$ paths that can be constructed with two $\langle -1,1\rangle$ steps, one $\langle 1,1\rangle$ step and two $\langle 0,1\rangle$ steps.

Matrix R. Next we derive expressions for each element of $\mathbf{R}$ in (5.46) using their probabilistic interpretation. Our analysis follows a similar approach to that above for the elements of the $\mathbf{G}$ matrix. We first introduce a useful definition.

Definition 5.2.5 *Let ζ_h denote the probability that the process reaches state $(i+1, j+h)$, $h \geq 0$, before the first return to any state of level i, given that the process started in state $(i+1, j)$.* □

From the probabilistic interpretation of $r_{j,k}$ and $r_h = r_{j,j+h}$ we see that

$$r_h = (\varphi\langle 1,0\rangle \zeta_h + \varphi\langle 1,1\rangle \zeta_{h-1}) \cdot \gamma \cdot \frac{[\mathbf{A}_1]_{jj}}{[\mathbf{A}_1]_{kk}}, \tag{5.55}$$

with γ the expected number of visits to state $(i+1, j+h)$ before the first return to any state of level i given state $(i+1, j+h)$ is reached at least once. Let us first consider γ. Say the process is in state $(i+1, j+h)$ for the first time. Then, the process will visit state $(i+1, j+h)$ for a second time, before it returns to level i, when it moves to $(i+2, j+h)$, starts a Bernoulli excursion of $2m$ steps ($m = 0, 1, \ldots$), and returns from state $(i+2, j+h)$ to $(i+1, j+h)$. With $\omega = \varphi\langle 1,0\rangle\varphi\langle -1,0\rangle \leq 1/4$, the probability that the process visits state $(i+1, j+h)$ for a second time is given by

$$\eta \;=\; \omega \sum_{m=0}^{\infty} \frac{1}{m+1}\binom{2m}{m}\omega^m \;=\; \tfrac{1}{2}(1-(1-4\omega)^{\frac{1}{2}}), \tag{5.56}$$

see, e.g., [202], on p. 561. Thus, $\gamma = 1 + \eta\gamma$ and so $\gamma = \frac{1}{1-\eta}$.

We have that $\zeta_0 = 1$ and for $s = 1, 2, \ldots$,

$$\zeta_s = \varphi\langle 0,1\rangle \zeta_{s-1} + \varphi\langle 1,0\rangle \sum_{j=0}^{s} G_{s-j}\zeta_j + \varphi\langle 1,1\rangle \sum_{j=0}^{s-1} G_{s-j-1}\zeta_j. \tag{5.57}$$

Rewriting (5.57) gives

$$\zeta_s = \frac{\varphi\langle 0,1\rangle \zeta_{s-1} + \varphi\langle 1,0\rangle \sum_{j=0}^{s-1} G_{s-j}\zeta_j + \varphi\langle 1,1\rangle \sum_{j=0}^{s-1} G_{s-j-1}\zeta_j}{1 - \varphi\langle 1,0\rangle G_0}. \tag{5.58}$$

Hypergeometric functions. Based on the results for the elements $r_{j,k}$ and $g_{j,k}$, we now derive explicit expressions for these fundamental matrix elements. More specifically, we express the elements of $\mathbf{G}$ and $\mathbf{R}$ in terms of the hypergeometric function defined as

$$F(a,b,c;z) \;=\; \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^n}{n!}, \tag{5.59}$$

with the gamma function $\Gamma(\cdot)$ defined as

$$\Gamma(z) \;=\; \int_0^{\infty} t^{z-1} e^{-t} dt, \qquad \mathrm{Re}(z) > 0. \tag{5.60}$$

We use the following result on the gamma function.

Lemma 5.2.6 *For $n = 1, 2, \ldots$ and $t = 1, 2, \ldots$,*

$$\Gamma(n + t/2) = \frac{\sqrt{\pi}}{2^{2(n-1)+t}} \frac{\Gamma(2n+t-1)}{\Gamma(n+t/2-1/2)}. \tag{5.61}$$

Proof. The case $t = 1$ can be found in Abramowitz and Stegun [25], on p. 255, and is given by, for $p = 1, 2, \ldots$,

$$\Gamma(p + 1/2) = \frac{\sqrt{\pi}}{2^{2p-1}} \frac{\Gamma(2p)}{\Gamma(p)}. \tag{5.62}$$

For $t = 2m$, $m = 1, 2, \ldots$, (5.61) reduces to

$$\Gamma(n+m) = \frac{\sqrt{\pi}}{2^{2(n+m-1)}} \frac{\Gamma(2(n+m)-1)}{\Gamma(n+m-1/2)}. \tag{5.63}$$

This gives

$$\Gamma(n+m-1/2) = \frac{\sqrt{\pi}}{2^{2(n+m-1)}} \frac{\Gamma(2(n+m)-1)}{\Gamma(n+m)} = \frac{\sqrt{\pi}}{2^{2(n+m-1)-1}} \frac{\Gamma(2(n+m-1))}{\Gamma(n+m-1)}, \tag{5.64}$$

which is equivalent to (5.62) for $p = m+n-1$. For the case $t = 2m+1$, $m = 0, 1, \ldots$, (5.61) reduces to

$$\Gamma(m+n+1/2) = \frac{\sqrt{\pi}}{2^{2(n+m)-1}} \frac{\Gamma(2(n+m))}{\Gamma(n+m)}, \tag{5.65}$$

which is equivalent to (5.62) for $p = m+n$. This completes the proof. □

We now consider the infinite series in (5.54),

$$\sum_{m=\max(u,s-1)}^{\infty} L_h(s,u,m)\cdot P_h(s,u,m). \tag{5.66}$$

In case $s-1\geq u$ we can write (5.66) as

$$\frac{\varphi\langle -1,1\rangle^s\varphi\langle 0,1\rangle^t\varphi\langle 1,1\rangle^u\varphi\langle 1,0\rangle^{s-1-u}}{s!\; t!\; u!}\sum_{m=0}^{\infty}\frac{\Gamma(2m+2s+t-1)}{\Gamma(m+s-u)}\frac{\omega^m}{m!}, \tag{5.67}$$

where $\omega=\varphi\langle 1,0\rangle\varphi\langle -1,0\rangle$. From Lemma 5.2.6 we get

$$\Gamma(2m+2s+t-1)=\frac{2^{2(m+s-1)+t}}{\sqrt{\pi}}\Gamma(m+s+\tfrac{t}{2})\Gamma(m+s+\tfrac{t-1}{2}), \tag{5.68}$$

and

$$\frac{2^{2s-2+t}}{\sqrt{\pi}}=\frac{\Gamma(2s+t-1)}{\Gamma(s+\frac{t}{2})\Gamma(s+\frac{t-1}{2})}. \tag{5.69}$$

Substituting (5.68) and (5.69) into the series in (5.67) yields

$$\begin{aligned}\sum_{m=0}^{\infty}\frac{\Gamma(2m+2s+t-1)}{\Gamma(m+s-u)}\frac{\omega^m}{m!} &= \frac{2^{2s-2+t}}{\sqrt{\pi}}\sum_{m=0}^{\infty}\frac{\Gamma(m+s+\frac{t}{2})\Gamma(m+s+\frac{t-1}{2})}{\Gamma(m+s-u)}\frac{(4\omega)^m}{m!}\\ &= \frac{(2s+t-2)!}{(s-u-1)!}\;F(s+\tfrac{t}{2},s+\tfrac{t-1}{2},s-u;4\omega),\end{aligned} \tag{5.70}$$

and, thus,

$$\begin{aligned}\sum_{m=s-1}^{\infty}L_h(s,u,m)\cdot P_h(s,u,m) &= \varphi\langle -1,1\rangle^s\varphi\langle 0,1\rangle^t\varphi\langle 1,1\rangle^u\varphi\langle 1,0\rangle^{s-1-u}\\ &\times\;\frac{(2s+t-2)!}{s!t!u!(s-u-1)!}\;F(s+\tfrac{t}{2},s+\tfrac{t-1}{2},s-u;4\omega).\end{aligned} \tag{5.71}$$

In a similar way, we find for the case that $u>s-1$,

$$\begin{aligned}\sum_{m=u}^{\infty}L_h(s,u,m)\cdot P_h(s,u,m) &= \varphi\langle -1,1\rangle^s\varphi\langle 0,1\rangle^t\varphi\langle 1,1\rangle^u\varphi\langle -1,0\rangle^{u+1-s}\\ &\times\;\frac{(2u+t)!}{s!t!u!(u-s+1)!}\;F(u+1+\tfrac{t}{2},u+1+\tfrac{t-1}{2},u-s+2;4\omega).\end{aligned} \tag{5.72}$$

The hypergeometric function $F(a,a+\frac{1}{2},c;z)$ emerging in the infinite series can be rewritten as

$$F(a+\tfrac{1}{2},a,c;z)=F(a,a+\tfrac{1}{2},c;z)=2^{c-1}\Gamma(c)z^{\frac{1}{2}-\frac{1}{2}c}(1-z)^{\frac{1}{2}c-a-\frac{1}{2}}P_{2a-c}^{1-c}((1-z)^{-\frac{1}{2}}), \tag{5.73}$$

with $P_\nu^\mu(\cdot)$ the Legendre function of the first kind (see [25], on p. 562). Using the recursion (see [25], on p. 334),

$$P_\nu^{\mu-1}(z)=\frac{P_{\nu+1}^{\mu}(z)-P_{\nu-1}^{\mu}(z)}{(2\nu+1)\sqrt{z^2-1}}, \tag{5.74}$$

we can recursively express the Legendre function $P_\nu^\mu(\cdot)$ in terms of Legendre polynomials $P_\nu^0(\cdot)$, i.e., $\mu = 0$. In turn, these Legendre polynomials have the following explicit form (see [25], on p. 775),

$$P_\nu^0(x) = \frac{1}{2^\nu} \sum_{m=0}^{\lfloor \frac{\nu}{2} \rfloor} (-1)^m \binom{\nu}{m} \binom{2(\nu-m)}{m} x^{\nu-2m}, \tag{5.75}$$

implying that all involved hypergeometric functions can be computed in a *finite* number of steps.

Application to the longest queue policy. The longest queue policy forms a particular instance of the introduced class. If (i, j) denotes the state that in equilibrium the difference $D = |L_1 - L_2|$ between the two queue lengths is i and the shortest queue $M = \min(L_1, L_2)$ is of length j, the QBD process describing the queue length process under the longest queue policy is described by the following matrices,

$$\mathbf{A}_0 = \begin{pmatrix} \lambda & & \\ & \lambda & \\ & & \ddots \end{pmatrix}, \quad \mathbf{B}_0 = \begin{pmatrix} 2\lambda & & & \\ \mu & 2\lambda & & \\ & \mu & 2\lambda & \\ & & \ddots & \ddots \end{pmatrix}, \tag{5.76}$$

and

$$\mathbf{A}_1 = \begin{pmatrix} \Delta & & \\ & \Delta & \\ & & \ddots \end{pmatrix}, \quad \mathbf{B}_1 = \begin{pmatrix} -2\lambda & & \\ & \Delta & \\ & & \ddots \end{pmatrix}, \quad \mathbf{A}_2 = \mathbf{B}_2 = \begin{pmatrix} \mu & \lambda & & \\ & \mu & \lambda & \\ & & \ddots & \ddots \end{pmatrix}, \tag{5.77}$$

where $\Delta = -(\mu + 2\lambda)$.

For the longest queue policy (with both $t = 0$ and $u = 0$), (5.71) significantly simplifies as shown below.

Proposition 5.2.7 *If $s = 0$ we have*

$$\sum_{m=0}^{\infty} L_h(0,0,m) \cdot P_h(0,0,m) = \tfrac{1}{2\varphi\langle 1,0\rangle}(1 - (1-4\omega)^{\frac{1}{2}}), \tag{5.78}$$

and if $s \geq 1$ we have

$$\sum_{m=s-1}^{\infty} L_h(s,0,m) \cdot P_h(s,0,m) = \varphi\langle -1,1\rangle^s \varphi\langle 1,0\rangle^{s-1} \frac{(2s-2)!}{s!(s-1)!}(1-4\omega)^{\frac{1}{2}-s}. \tag{5.79}$$

Proof. (5.78) follows directly from (5.56). From (5.71) it follows that

$$\sum_{m=s-1}^{\infty} L_h(s,0,m) \cdot P_h(s,0,m) = \varphi\langle -1,1\rangle^s \varphi\langle 1,0\rangle^{s-1} \frac{(2s-2)!}{s!(s-1)!} F(s, s-\tfrac{1}{2}, s; 4\omega). \tag{5.80}$$

The fact that $F(s, s-\frac{1}{2}, s; 4\omega) = F(s-\frac{1}{2}, s, s; 4\omega)$ together with the identity $F(a,b,b;z) = (1-z)^{-a}$ (see [25], on p. 556) then completes the proof. □

The results presented here may be used to establish asymptotic results on the powers of the **R** matrix, which can be used together with the corresponding matrix element expressions to obtain explicit large-deviation decay rates. Furthermore, in [P2; P3] we extend the class of models for which an explicit expression for the fundamental matrices can be obtained in several directions (more general transition rates, finite dimension of the phase and more general boundary states). Some classical queueing models, relevant for production environments, that fit within this extended class are the multi-machine lowest inventory policy, the two-class Markovian model with non-preemptive priority [125], a hybrid make-to-order/make-to-stock system [27], a two-machine re-entrant line [28] and many more. Finally, [P3] presents for a subclass a recursive algorithm for calculating the elements of the matrices in an efficient way.

Bibliography

Self-references

[P1] Adan, I.J.B.F., Leeuwaarden, J.S.H. van, Winands, E.M.M., (2006). *On the application of Rouché's theorem in queueing theory* (Operations Research Letters, vol. 34, no. 3, pp. 355-360).

[P2] Leeuwaarden, J.S.H. van, Squillante, M.S., Winands, E.M.M., (2007). *QBD processes, lattice path counting, and hypergeometric functions* (Report, Eindhoven University of Technology).

[P3] Leeuwaarden, J.S.H. van, Winands, E.M.M., (2006). *Quasi-birth-and-death processes with an explicit rate matrix* (Stochastic Models, vol. 22, no. 1, pp. 77-98).

[P4] Mei, R.D. van der, Winands, E.M.M., (2006). *Mean value analysis for polling systems in heavy traffic* (Proceedings of Valuetools, ACM Press).

[P5] Mei, R.D. van der, Winands, E.M.M., (2007). *Heavy traffic analysis of polling models by mean value analysis* (To appear in Performance Evaluation - Special issue for best papers of Valuetools 2006).

[P6] Mei, R.D. van der, Winands, E.M.M., (2007). *Polling models with renewal arrivals: a new method to derive heavy-traffic asymptotics* (To appear in Performance Evaluation).

[P7] Mei, R.D. van der, Winands, E.M.M., (2007). *A note on polling models with renewal arrivals and nonzero switch-over times* (Report, Eindhoven University of Technology).

[P8] Vuuren, M. van, Winands, E.M.M., (2007). *Iterative approximation of k-limited polling systems* (Queueing Systems, vol. 55, no. 3, pp. 161-178).

[P9] Wierman, A.C., Winands, E.M.M., Boxma, O.J., (2007). *Scheduling in polling systems* (To appear in Performance Evaluation).

[P10] Winands, E.M.M., (2007). *On polling systems with large setups* (To appear in Operations Research Letters).

[P11] Winands, E.M.M., (2007). *Branching-type polling systems with large setups* (Report, Eindhoven University of Technology).

[P12] Winands, E.M.M., (2007). *On variances in polling systems* (Report, Eindhoven University of Technology).

[P13] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, (2005). *The stochastic economic lot scheduling problem: a survey* (Invited review for European Journal of Operational Research, under revision).

[P14] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, (2005). *A two-queue model with alternating limited service and state-dependent setups* (Proceedings of Analysis of Manufacturing Systems, pp. 200-208).

[P15] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, (2006). *Mean value analysis for polling systems* (Queueing Systems, vol. 54, no. 1, pp. 45-54).

[P16] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, (2007). *MVA for polling systems: an efficient approach* (Proceedings of Analysis of Manufacturing Systems, pp. 147-153).

[P17] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, (2007). *A unifying MVA framework for polling systems: exact results and asymptotics* (Report, Eindhoven University of Technology).

[P18] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, Down, D.G., (2007). *A state-dependent polling model with k-limited service* (To appear in Probability in the Engineering and Informational Sciences).

[P19] Winands, E.M.M., Denteneer, T.J.J., Resing, J.A.C., Rietman, R., (2004). *A finite-source feedback queueing network as a model for the IEEE 802.11 distributed coordination function* (Proceedings of European Wireless 2004, pp. 551-557).

[P20] Winands, E.M.M., Denteneer, T.J.J., Resing, J.A.C., Rietman, R., (2005). *A finite-source queueing model for the IEEE 802.11 DCF* (European Transactions on Telecommunications - Special issue for best papers of European Wireless 2004, vol. 16, no. 1, pp. 77-89).

[P21] Winands, E.M.M., Kok, A.G. de, Timpe, C., (2007). *Case study of a batch-production/inventory system* (Report, BASF).

[P22] Winands, E.M.M., Kreuk, A.C.C. de, Vissers, J.M.H., (2005). *Master scheduling of medical specialists* (Health Operations Management, J. Vissers and R. Beech (eds.), Routledge, London, pp. 184-201).

[P23] Winands, E.M.M., Wieland, J., Sanders, B., (2006). *Dynamic half-rate connections in GSM* (AEÜ - International Journal of Electronics and Communications, vol. 60, no. 7, pp. 504-512).

References

[24] Abate, J., Whitt, W., (1992). *Numerical inversion of probability generating functions* (Operations Research Letters, vol. 12, no. 4, pp. 245-251).

[25] Abramowitz, M., Stegun, I.A., (1964). *Handbook of Mathematical Functions* (US Government Printing Office, Washington D.C.).

[26] Adan, I.J.B.F., Houtum, G.J. van, Wal, J. van der, (1997). *The symmetric longest queue system* (Stochastic Models, vol. 13, pp. 105-120).

[27] Adan, I.J.B.F., Wal, J. van der, (1998). *Combining make to order and make to stock* (OR Spektrum, vol. 20, pp. 73-81).

[28] Adan, I.J.B.F., Weiss, G., (2006). *Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy* (Queueing Systems, vol. 54, pp. 169-183).

[29] Altiok, T., Shiue, G.A., (1994). *Single-stage, multi-product production/inventory systems with backorders* (IIE Transactions, vol. 26, no. 2, pp. 52-61).

[30] Altiok, T., Shiue, G.A., (1995). *Single-stage, multi-product production/inventory systems with lost sales* (Naval Research Logistics, vol. 42, pp. 889-913).

[31] Altiok, T., Shiue, G.A., (2000). *Pull-type manufacturing systems with multiple product types* (IIE Transactions, vol. 32, no. 2, pp. 115-124).

[32] Altman, E., Blanc, H., Khamisy, A., Yechiali, U., (1994). *Gated-type polling systems with walking and switch-in times* (Stochastic Models, vol. 10, pp. 741-764).

[33] Altman, E., Konstantopoulos, P., Liu, Z., (1992). *Stability, monotonicity and invariant quantities in general polling systems* (Queueing Systems, vol. 11, pp. 35-57).

[34] Altman, E., Yechiali, U., (1994). *Polling in a closed network* (Probability in the Engineering and Informational Sciences, vol. 8, no. 3, pp. 327-343).

[35] Amrony, R., Yechiali, U., (1999). *Polling systems with permanent and transient customers* (Stochastic Models, vol. 15, no. 3, pp. 395-427).

[36] Anupindi, R., Tayur, S., (1998). *Managing stochastic multiproduct systems: model, measures, and analysis* (Operations Research, vol. 46, no. 3S, pp. 98-111).

[37] Asmussen, S., Koole, G., (1993). *Marked point processes as limits of Markovian arrival streams* (Journal of Applied Probability, vol. 30, pp. 365-372).

[38] Bertsekas, D., Gallager, R., (1987). *Data Networks* (Prentice-Hall, New Jersey).

[39] Bertsimas, D., Mourtzinou, G., (1997). *Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws* (Operations Research, vol. 45, pp. 470-487).

[40] Bertsimas, D., Mourtzinou, G., (1999). *Decomposition results for general polling systems and their applications* (Queueing Systems, vol. 31, pp. 295-316).

[41] Blanc, J.P.C., (1992). *An algorithmic solution of polling models with limited service disciplines* (IEEE Transactions on Communications, vol. 40, no. 7, pp. 1152-1155).

[42] Borst, S.C., Boxma, O.J., (1997). *Polling models with and without switchover times* (Operations Research, vol. 45, no. 4, pp. 536-543).

[43] Borst, S.C., Boxma, O.J., Levy, H., (1995). *The use of service limits for efficient operation of multistation single-medium communication systems* (IEEE/ACM Transactions on Networking, vol. 3, no. 5, pp. 602-612).

[44] Bourland, K.E., (1994). *Production planning and control links and the stochastic economic lot scheduling problem* (Working paper no. 299, The Amos Tuck School of Business Administration).

[45] Bourland, K.E., Yano, C.A., (1994). *The strategic use of capacity slack in the economic lot scheduling problem with random demand* (Management Science, vol. 40, no. 12, pp. 1690-1704).

[46] Boxma, O.J., (1985). *Two symmetric queues with alternating service and switching times* (In Performance '84, E. Gelenbe (ed.), North-Holland, Amsterdam, pp. 409-431).

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

[48] Boxma, O.J., Down, D.G., (1997). *Dynamic server assignment in a two-queue model* (European Journal of Operational Research, vol. 103, pp. 101-115).

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

[50] Boxma, O.J., Levy, H., Weststrate, J., (1991). *Efficient visit frequencies for polling tables: minimization of waiting cost* (Queueing Systems, vol. 9, no. 1-2, pp. 133-162).

[51] Boxma, O.J., Levy, H., Weststrate, J., (1993). *Efficient visit orders for polling systems* (Performance Evaluation, vol. 18, pp. 103-123).

[52] Boxma, O.J., Levy, H., Yechiali, U., (1992). *Cyclic reservation schemes for efficient operation of multiple-queue single-server systems* (Annals of Operations Research, vol. 35, pp. 187-208).

[53] Brander, P., (2005). *Inventory Control and Scheduling Problems in a Single-Machine Multi-Item System* (Ph.D. Thesis, Lulea University of Technology).

[54] Brander, P., Forsberg, R., (2006). *Determination of safety stocks for cyclic schedules with stochastic demands* (International Journal of Production Economics, vol. 104, pp. 271-295).

[55] Brander, P., Léven, E., Segerstedt, A., (2005). *Lot sizes in a capacity constrained facility - a simulation study of stationary stochastic demand* (International Journal of Production Economics, vol. 93-94, pp. 375-386).

[56] Broek, M.S. van den, (2004). *Traffic Signals: Optimizing and Analyzing Traffic Control Systems* (Master's thesis, Eindhoven University of Technology).

[57] Broek, M.S. van den, Leeuwaarden, J.S.H. van, Adan, I.J.B.F., Boxma, O.J., (2006). *Bounds and approximations for the fixed-cyle traffic-light queue* (Transportation Science, vol. 40, no. 4, pp. 484-496).

[58] Bruin, J., (2007). *Cyclic multi-item production systems* (Proceedings of Analysis of Manufacturing Systems, pp. 99-104).

[59] Chang, K.C., Sandhu, D., (1992). *Pseudo-conservation laws in cyclic-service systems with a class of limited service policies* (Annals of Operations Research, vol. 35, pp. 209-229).

[60] Chang, K.C., Sandhu, D., (1992). *Mean waiting time approximations in cyclic-service systems with exhaustive limited service policy* (Performance Evaluation, vol. 15, no. 1, pp. 21-40).

[61] Chang, W., Down, D.G., (2002). *Exact asymptotics for k_i-limited exponential polling models* (Queueing Systems, vol. 42, no. 4, pp. 401-419).

[62] Chang, W., Down, D.G., (2007). *Polling models under limited service policies: sharp asymptotics* (Stochastic Models, vol. 23, pp. 129-147).

[63] Charzinski, J., Renger, T., Tangemann, M., (1994). *Simulative comparison of the waiting time distributions in cyclic polling systems with different service strategies* (Proceedings of the 14th International Teletraffic Congress, pp. 719-728).

[64] Chaudhry, M.L., (1994). *QROOT Software Package* (A&A Publications, Kingston).

[65] Choudhury, G.L., Whitt, W., (1996). *Computing distributions and moments in polling models by numerical transform inversion* (Performance Evaluation, vol. 25, no. 4, pp. 267-292).

[66] Coffman, E.G., Puhalskii, A.A., Reiman, M.I., (1995). *Polling systems with zero switch-over times: a heavy-traffic principle* (The Annals of Applied Probability, vol. 5, pp. 681-719).

[67] Coffman, E.G., Puhalskii, A.A., Reiman, M.I., (1998). *Polling systems in heavy-traffic: a Bessel process limit* (Mathematics of Operations Research, vol. 23, pp. 257-304).

[68] Cohen, J.W., (1969). *The Single Server Queue* (North-Holland Publishing Company, Amsterdam).

[69] Cohen, J.W., (1987). *A two-queue, one-server model with priority for the longer queue* (Queueing Systems, vol. 2, no. 3, pp. 261-283).

[70] Cohen, J.W., Boxma, O.J., (1981). *The M/G/1 queue with alternating service formulated as a Riemann-Hilbert problem* (In Performance '81, F.J. Kylstra (ed.), North-Holland, Amsterdam, pp. 181-189).

[71] Cooper, R.B., (1970). *Queues served in cyclic order: waiting times* (The Bell System Technical Journal, vol. 49, pp. 399-413).

[72] Cooper, R.B., Murray, G., (1969). *Queues served in cyclic order* (The Bell System Technical Journal, vol. 48, pp. 675-689).

[73] Cooper, R.B., Niu, S.-C., Srinivasan, M.M., (1996). *A decomposition theorem for polling models: the switchover times are effectively additive* (Operations Research, vol. 44, pp. 629-633).

[74] Cooper, R.B., Niu, S.-C., Srinivasan, M.M., (1998). *When does forced idle time improve performance in polling models?* (Management Science, vol. 44, pp. 1079-1086).

[75] Dai, J.G., (1995). *On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models* (Annals of Applied Probability, vol. 5, pp. 49-77).

[76] Dai, J.G., Meyn, S.P., (1995). *Stability and convergence of moments for multiclass queueing networks via fluid models* (IEEE Transactions on Automatic Control, vol. 40, pp. 1889-1904).

[77] Dallery, Y., David, R., Xie, X., (1989). *Approximate analysis of transfer lines with unreliable machines and finite buffers* (IEEE Transactions on Automatic Control, vol. 34, no. 9, pp. 943-953).

[78] Dellaert, N.P., (1989). *Production To Order: Models and Rules for Production Planning* (Lecture Notes in Economics and Mathematical Systems 333, Springer Verlag, Berlin).

[79] Doshi, B.T., (1986). *Queueing systems with vacations - a survey* (Queueing Systems, vol. 1, no. 1, pp. 29-66).

[80] Doshi, B.T., (1990). *Single server queues with vacations* (In Stochastic Analysis of Computer and Communication Systems, H. Takagi (ed.), North-Holland, Amsterdam, pp. 217-265).

[81] Down, D., (1998). *On the stability of polling models with multiple servers* (Journal of Applied Probability, vol. 35, pp. 925-935).

[82] Duenyas, I., (1994). *The limitations of suboptimal policies* (Interfaces, vol. 24, no. 5, pp. 77-84).

[83] Eisenberg, M., (1971). *Two queues with changeover times* (Operations Research, vol. 19, no. 3, pp. 386-401).

[84] Eisenberg, M., (1979). *Two queues with alternating service* (SIAM Journal on Applied Mathematics, vol. 36, no. 2, pp. 287-303).

[85] Eisenstein, D.D., (2005). *Recovering cyclic schedules using dynamic produce-up-to policies* (Operations Research, vol. 53, no. 4, pp. 675-688).

[86] Elmaghraby, S.E., (1978). *The economic lot scheduling problem (ELSP): review and extensions* (Management Science, vol. 24, no. 6, pp. 587-598).

[87] Erkip, N., Güllü, R., Kocabiyikoglu, A., (2000). *A quasi-birth-and-death model to evaluate fixed cycle time policies for stochastic multi-item production/inventory problem* (Proceedings of MSOM conference).

[88] Everitt, D., (1989). *A note on the pseudoconservation laws for cyclic service systems with limited service disciplines* (IEEE Transactions on Communications, vol. 37, no. 7, pp. 781-783).

[89] Federgruen, A., Katalan, Z., (1994). *Approximating queue size and waiting time distributions in general polling systems* (Queueing Systems, vol. 18, pp. 353-386).

[90] Federgruen, A., Katalan, Z., (1996). *The impact of setup times on the performance of multi-class service and production systems* (Operations Research, vol. 44, pp. 989-1001).

[91] Federgruen, A., Katalan, Z., (1996). *The stochastic economic lot scheduling problem: cyclical base-stock policies with idle times* (Management Science, vol. 42, no. 6, pp. 783-796).

[92] Federgruen, A., Katalan, Z., (1996). *Customer waiting-time distributions under base-stock policies in single-facility multi-item production systems* (Naval Research Logistics, vol. 43, pp. 533-548).

[93] Federgruen, A., Katalan, Z., (1998). *Determining production schedules under base-stock policies in single facility multi-item production systems* (Operations Research, vol. 46, no. 6, pp. 883-898).

[94] Federgruen, A., Katalan, Z., (1999). *The impact of adding a make-to-order item to a make-to-stock production system* (Management Science, vol. 45, no. 7, pp. 980-994).

[95] Ferguson, M.J., (1986). *Computation of the variance of the waiting time for token rings* (IEEE Journal on Selected Areas in Communications, vol. SAC-4, pp. 775-782).

[96] Ferguson, M.J., Aminetzah, Y., (1985). *Exact results for nonsymmetric token ring systems* (IEEE Transactions on Communications, vol. COM-33, pp. 223-231).

[97] Flatto, L., (1989). *The longer queue model* (Probability in the Engineering and Informational Sciences, vol. 3, pp. 537-559).

[98] Fournier, L., Rosberg, Z., (1991). *Expected waiting times in polling systems under priority disciplines* (Queueing Systems, vol. 9, pp. 419-440).

[99] Franken, P., Koenig, D., Arndt, W., Schmidt, F., (1982). *Queues and Point Processes* (John Wiley, New York).

[100] Fransoo, J.C., (1992). *Demand management and production control in process industries* (International Journal of Operations and Production Management, vol. 12, no. 7/8, pp. 187-196).

[101] Fransoo, J.C., (1993). *Production Control and Demand Management in Capacitated Flow Process Industries* (Ph.D. Thesis, Eindhoven University of Technology).

[102] Fransoo, J.C., Sridharan, V., Bertrand, J.W.M., (1995). *A hierarchical approach for capacity coordination in multiple products single-machine production systems with stationary stochastic demands* (European Journal of Operational Research, vol. 86, no. 1, pp. 57-72).

[103] Fricker, C., Jaibi, R., (1994). *Monotonicity and stability of periodic polling models* (Queueing Systems, vol. 15, pp. 211-238).

[104] Fuhrmann, S.W., (1981). *Performance analysis of a class of cyclic schedules* (Bell Laboratories Technical Memorandum 81-59531-1).

[105] Fuhrmann, S.W., (1985). *Symmetric queues served in cyclic order* (Operations Research Letters, vol. 4, no. 3, pp. 139-144).

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

[107] Gallego, G., (1990). *Scheduling the production of several items with random demands in a single facility* (Management Science, vol. 36, no. 12, pp. 1579-1592).

[108] Gallego, G., (1994). *When is a base stock policy optimal in recovering disrupted cyclic schedules?* (Naval Research Logistics, vol. 41, pp. 317-333).

[109] Gallego, G., Moon, I., (1993). *The distribution free newsboy problem: review and extensions* (The Journal of the Operational Research Society, vol. 44, no. 8, pp. 825-834).

[110] Gascon, A., Leachman, R.C., Lefrançois, P., (1994). *Multi-item, single-machine scheduling problem with stochastic demands: a comparison of heuristics* (International Journal of Production Research, vol. 32, pp. 583-596).

[111] Gerchak, Y., Zhang, Z., (1994). *The cheaper/faster-yet-more-expensive phenomenon: are Zangwill's 'paradoxes' indeed paradoxical?* (Interfaces, vol. 24, no. 5, pp. 84-87).

[112] Gershwin, S.B., Burman, M.H., (2000). *A decomposition method for analyzing inhomogeneous assembly/disassembly systems* (Annals of Operation Research, vol. 93, pp. 91-115).

[113] Giezenaar, R.B.L.M., (1997). *Ontwerp voor een Productie- en Voorraadstrategie voor de Productieplant Laurox bij AKZO Nobel Chemicals te Deventer* (Master's thesis, University of Twente, in Dutch).

[114] Grasman, S.E., Olsen, T.L., Birge, J.R., (2004). *Setting basestock levels in multiproduct systems with setups and random yield* (Working paper, Olin School of Business, St. Louis).

[115] Graves, S.C., (1980). *The multi-product production cycling problem* (AIIE Transactions, vol. 12, pp. 233-240).

[116] Groenendijk, W.P., (1990). *Conservation Laws in Polling Systems* (Ph.D. Thesis, University of Utrecht).

[117] Günalay, Y., Gupta, D., (1997). *A polling system with a patient server and state-dependent setup times* (IIE Transactions, vol. 29, pp. 469-480).

[118] Gupta, D., Srinivasan, M.M., (1996). *Polling systems with state-dependent setup times* (Queueing Systems, vol. 22, pp. 403-423).

[119] Ha, A.Y., (1997). *Optimal dynamic scheduling policy for a make-to-stock production system* (Operations Research, vol. 45, pp. 42-53).

[120] Hirayama, T., Hong, S.J., Krunz, M., (2004). *A new approach to analysis of polling systems* (Queueing Systems, vol. 48, no. 1-2, pp. 135-158).

[121] Hsu, W.L., (1983). *On the general feasibility test of scheduling lot sizes for several products on one machine* (Management Science, vol. 29, no. 1, pp. 93-105).

[122] Ibe, O.C., (1990). *Analysis of polling systems with mixed service disciplines* (Stochastic Models, vol. 6, pp. 667-689).

[123] Ibe, O.C., Cheng, X., (1988). *Stability conditions for multiqueue systems with cyclic service* (IEEE Transactions Automatic Control, vol. 33, no. 1, pp. 102-103).

[124] Inman, R.R., (1999). *Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems* (Production and Operations Management, vol. 8, no. 4, pp. 409-432).

[125] Jaiswal, N.K., (1968). *Priority Queues* (Academic Press, London).

[126] Johnson, M.A., (1993). *An empirical study of queueing approximations based on phase-type distributions* (Stochastic Models, vol. 9, no. 4, pp. 531-561).

[127] Karmarkar, U.S., Yoo, J., (1994). *Stochastic dynamic product cycling problem* (European Journal of Operational Research, vol. 73, pp. 360-373).

[128] Kat, B., Avsar, Z.M., (2005). *Heuristics for dynamic scheduling of multi-class base-stock controlled systems* (Proceedings of Analysis of Manufacturing Systems, pp. 217-224).

[129] Keilson, J., Servi, L.D., (1986). *Oscillating random walk models for GI/G/I vacation systems with Bernoulli schedules* (Journal of Applied Probability, vol. 23, pp. 790-802).

[130] Keilson, J., Servi, L.D., (1990). *The distributional form of Little's law and the Fuhrmann-Cooper decomposition* (Operations Research Letters, vol. 9, no. 4, pp. 239-247).

[131] Kitaev, M. Yu., (1993). *The $M/G/1$ processor-sharing model: transient behavior* (Queueing Systems, vol. 14, pp. 239-273).

[132] Konheim, A.G., Levy, H., Srinivasan, M.M., (1994). *Descendant set: an efficient approach for the analysis of polling systems* (IEEE Transactions on Communications, vol. 42, no. 2/3/4, pp. 1245-1253).

[133] Konheim, A.G., Meister, B., (1974). *Waiting lines and times in a system with polling* (Journal of the Association for Computing Machinery, vol. 21, no. 3, pp. 470-490).

[134] Krieg, G.N., Kuhn, H., (2002). *A decomposition method for multi-product kanban systems with setup times and lost sales* (IIE Transactions, vol. 34, no. 7, pp. 613-625).

[135] Krieg, G.N., Kuhn, H., (2004). *Analysis of multi-product kanban systems with state-dependent setups and lost sales* (Annals of Operations Research, no. 125, pp. 141-166).

[136] Kroese, D.P., (1997). *Heavy traffic analysis for continuous polling models* (Journal of Applied Probability, vol. 34, pp. 720-732).

[137] Kudoh, S., Takagi, H., Hashida, O., (2000). *Second moments of the waiting time in symmetric polling systems* (Journal of the Operations Research Society of Japan, vol. 43, no. 2, pp. 306-316).

[138] Kühn, P.J., (1979). *Multiqueue systems with nonexhaustive cyclic service* (Bell System Technical Journal, vol. 58, no. 3, pp. 671-698).

[139] Lang, M., Bosch, M., (1991). *Performance analysis of finite capacity polling systems with limited-M service* (Proceedings of the 13th International Teletraffic Congress, pp. 731-735).

[140] Latouche, G., Ramaswami, V., (1993). *A logarithmic reduction algorithm for quasi-birth-and-death processes* (Journal of Applied Probability, vol. 30, pp. 650-674).

[141] Latouche, G., Ramaswami, V., (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling* (SIAM, Philadelphia).

[142] Lavenberg, S.S., (1983). *Computer Performance Modeling Handbook* (Academic Press, London).

[143] Leachman, R.C., Gascon, A., (1988). *A heuristic policy for multi-item, single-machine production systems with time-varying stochastic demands* (Management Science, vol. 34, no. 3, pp. 377-390).

[144] Leachman, R.C., Xiong, Z.K., Gascon, A., Park, K., (1991). *An improvement to the dynamic cycle lengths heuristic for scheduling the multi-item, single-machine* (Management Science, vol. 37, no. 9, 1201-1205).

[145] Lee, D.-S., (1996). *A two-queue model with exhaustive and limited service disciplines* (Stochastic Models, vol. 12, no. 2, pp. 285-305).

[146] Lee, D.-S., Sengupta, B., (1992). *An approximate analysis of a cyclic server queue with limited service and reservations* (Queueing Systems, vol. 11, pp. 153-178).

[147] Leung, K.K., (1991). *Cyclic-service systems with probabilistically-limited service* (IEEE Journal on Selected Areas in Communications, vol. 9, no. 2, pp. 185-193).

[148] Levy, H., (1988). *Optimization of polling systems: the fractional exhaustive service method* (Report, Tel-Aviv University).

[149] Levy, H., (1989). *Analysis of cyclic polling systems with binomial gated service* (In Performance of Distributed and Parallel Systems, T. Hasegawa, H. Takagi and Y. Takahashi (eds.), North-Holland, Amsterdam, pp. 127-139).

[150] Levy, H., (1989). *Delay computation and dynamic behavior of non-symmetric polling systems* (Performance Evaluation, vol. 10, no. 1, pp. 35-51).

[151] Levy, H., Sidi, M., (1990). *Polling systems: applications, modeling and optimization* (IEEE Transactions on Communications, vol. COM-38, no. 10, pp. 1750-1760).

[152] Levy, H., Sidi, M., Boxma, O.J., (1990). *Dominance relations in polling systems* (Queueing Systems, vol. 6, no. 2, pp. 155-171).

[153] Levy, Y., (1985). *A class of scheduling policies for real-time processors with switching system applications* (Proceedings of the 11th International Teletraffic Congress, pp. 760-766).

[154] Little, J.D.C., (1961). *A proof of the queueing formula $L = \lambda W$* (Operations Research, vol. 9, pp. 383-387).

[155] Mack, C., (1957). *The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable* (Journal of the Royal Statistical Society Series B, vol. 19, no. 1, pp. 173-178).

[156] Mack, C., Murphy, T., Webb, N.L., (1957). *The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants* (Journal of the Royal Statistical Society Series B, vol. 19, no. 1, pp. 166-172).

[157] Markowitz, D.M., Reiman, M.I., Wein, L.M., (2000). *The stochastic economic lot scheduling problem: heavy traffic analysis of dynamic cyclic policies* (Operations Research, vol. 48, no. 1, pp. 136-154).

[158] Markowitz, D.M., Wein, L.M., (2001). *Heavy traffic analysis of dynamic cyclic policies: a unified treatment of the single machine scheduling problem* (Operations Research, vol. 49, no. 2, pp. 246-270).

[159] McIntyre, B., (1994). *A comment on Zangwill's 'The limits of Japanese production theory'* (Interfaces, vol. 24, no. 5, pp. 87-89).

[160] Mei, R.D. van der, (1995). *Polling Systems and the Power-Series Algorithm* (Ph.D. Thesis, University of Tilburg).

[161] Mei, R.D. van der, (1999). *Polling systems in heavy traffic: higher moments of the delay* (Queueing Systems, vol. 31, pp. 265-294).

[162] Mei, R.D. van der, (1999). *Distributions of the delay in polling systems in heavy traffic* (Performance Evaluation, vol. 38, pp. 133-148).

[163] Mei, R.D. van der, (1999). *Delay in polling systems with large switch-over times* (Journal of Applied Probability, vol. 36, pp. 232-243).

[164] Mei, R.D. van der, (2000). *Polling systems with switch-over times under heavy load: moments of the delay* (Queueing Systems, vol. 36, pp. 381-404).

[165] Mei, R.D. van der, (2006). *Towards a unifying theory on branching-type polling models in heavy traffic* (Report, Vrije Universiteit).

[166] Mei, R.D. van der, Levy, H., (1997). *Polling systems in heavy traffic: exhaustiveness of the service policies* (Queueing Systems, vol. 27, pp. 227-250).

[167] Mei, R.D. van der, Levy, H., (1998). *Expected delay in polling systems in heavy traffic* (Advances in Applied Probability, vol. 30, pp. 586-602).

[168] Menich, R., Serfozo, R.F., (1991). *Optimality of routing and servicing in dependent parallel processing systems* (Queueing Systems, vol. 9, no. 4, pp. 403-418).

[169] Naoumov, V.A., Krieger, U.R., Wagner, D., (1997). *Analysis of a multiserver delay-loss system with a general Markovian arrival process* (Matrix-Analytic Methods in Stochastic Models, A.S. Alfa and S.R. Chakravarthy (eds.), Marcel Dekker, New York, pp. 43-66).

[170] Neuts, M.F., (1981). *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach* (The Johns Hopkins Press, Baltimore).

[171] Niño-Mora, J., (2006). *Marginal productivity index policies for scheduling a multiclass delay-/loss-sensitive queue* (Queueing Systems, vol. 54, no. 4, pp. 281-312).

[172] Nyen, P.L.M. van, (2005). *The Integrated Control of Production-Inventory Systems* (Ph.D. Thesis, Eindhoven University of Technology).

[173] Olsen, T.L., (2001). *Limit theorems for polling models with increasing setups* (Probability in the Engineering and Informational Sciences, vol. 15, pp. 35-55).

[174] Olsen, T.L., Mei, R.D. van der, (2003). *Periodic polling systems in heavy-traffic: distribution of the delay* (Journal of Applied Probability, vol. 40, pp. 305-326).

[175] Olsen, T.L., Mei, R.D. van der, (2005). *Periodic polling systems in heavy-traffic: renewal arrivals* (Operations Research Letters, vol. 33, pp. 17-25).

[176] Ozawa, T., (1990). *Alternating service queues with mixed exhaustive and K-limited services* (Performance Evaluation, vol. 11, pp. 165-175).

[177] Ozawa, T., (1997). *Waiting time distribution in a two-queue model with mixed exhaustive and gated-type K-limited services* (Proceedings of International Conference on the Performance and Management of Complex Communication Networks, pp. 231-250).

[178] Paternina-Arboleda, C.D., Das, T.K., (2005). *A multi-agent reinforcement learning approach to obtaining dynamic control policies for stochastic lot scheduling problem* (Simulation Modelling Practice and Theory, vol. 13, pp. 389-406).

[179] Qiu, J., Loulou, R., (1995). *Multiproduct production/inventory control under random demands* (IEEE Transactions on Automatic Control, vol. 40, no. 2, pp. 350-356).

[180] Quine, M.P., (1972). *The multitype Galton-Watson process with ρ near 1* (Advances in Applied Probability, vol. 4, pp. 429-452).

[181] Reiman, M.I., Wein, L.M., (1998). *Dynamic scheduling of a two-class queue with setups* (Operations Research, vol. 46, pp. 532-547).

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

[183] Ross, S.M., (1983). *Stochastic Processes* (John Wiley, New York).

[184] Rubin, I., De Moraes, L.F.M., (1983). *Message delay analysis for polling and token multiple-access schemes for local communication networks* (IEEE Journal on Selected Areas in Communications, vol. SAC-l, no. 5, pp. 935-947).

[185] Ruigt, P., (2006). *Design of a Planning Control Model in an Oleochemical Production Environment* (Master's thesis, Eindhoven University of Technology).

[186] Salomon, M., (1990). *Determining Lotsizing Models for Production Planning* (Ph.D. Dissertation, Erasmus University Rotterdam).

[187] Sarkar, D., Zangwill, W.I., (1989). *Expected waiting time for nonsymmetric cyclic queueing systems - exact results and applications* (Management Science, vol. 35, pp. 1463-1474).

[188] Sarkar, D., Zangwill, W.I., (1991). *Variance effects in cyclic production systems* (Management Science, vol. 37, pp. 444-453).

[189] Schassberger, R.S., (1970). *On the waiting time in the queueing system $GI/G/1$* (The Annals of Mathematical Statistics, vol. 41, no. 1, pp. 182-187).

[190] Schassberger, R.S., (1973). *Warteschlangen* (Springer-Verlag, Berlin).

[191] Schmidt, E., Dada, M., Ward, J., Adams, D., (2001). *Using cyclic planning to manage capacity at ALCOA* (Interfaces, vol. 31, pp. 16-27).

[192] Shimogawa, S., Takahashi, Y., (1992). *A note on the pseudo-conservation law for a multi-queue with local priority* (Queueing Systems, vol. 11, no. 1-2, pp. 145-151).

[193] Singh, M.P., Srinivasan, M.M., (2002). *Exact analysis of the state dependent polling model* (Queueing Systems, vol. 41, pp. 371-399).

[194] Smits, S.R., Wagner, M., Kok, A.G. de, (2004). *Determination of an order-up-to policy in the stochastic economic lot scheduling model* (International Journal of Production Economics, vol. 90, pp. 377-389).

[195] Sox, C.R., Jackson, P.L., Bowman, A., Muckstadt, J.A., (1999). *A review of the stochastic lot scheduling problem* (International Journal of Production Economics, vol. 62, pp. 181-200).

[196] Sox, C.R., Muckstadt, J.A., (1997). *Optimization-based planning for the stochastic lot scheduling problem* (IIE Transactions, vol. 29, no. 5, pp. 349-357).

[197] Srinivasan, M.M., Levy, H., Konheim, A.G., (1996). *The individual station technique for the analysis of polling systems* (Naval Research Logistics, vol. 43, no. 1, pp. 79-101).

[198] Srinivasan, M.M., Niu, S.-C., Cooper, R.B., (1995). *Relating polling models with zero and nonzero zwitchover times* (Queueing Systems, vol. 19, pp. 149-168).

[199] Stoop, P.P.M., (1996). *Performance Management in Manufacturing, A Method for Short Term Performance Evaluation and Diagnosis* (Ph.D. Thesis, Eindhoven University of Technology).

[200] Swartz, G.B., (1980). *Polling in a loop system* (Journal of the Association for Computing Machinery, vol. 27, no. 1, pp. 42-59).

[201] Takács, L., (1968). *Two queues attended by a single server* (Operations Research, vol. 16, pp. 639-650).

[202] Takács, L., (1991). *A Bernoulli excursion and its various applications* (Advances in Applied Probability, vol. 23, no. 3, pp. 557-585).

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

[204] Takagi, H., (1991). *Queueing Analysis: A Foundation of Performance Evaluation, Vacation and Priority Systems, part 1* (North-Holland, Amsterdam).

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

[206] Takagi, H., (2000). *Analysis and application of polling models* (In Performance Evaluation: Origins and Directions, G. Haring, C. Lindemann and M. Reiser (eds.), Lecture Notes in Computer Science, vol. 1769, Springer, Berlin, pp. 423-442).

[207] Takahashi, Y., Kumar, B.K., (1995). *Pseudo-conservation law for a priority polling system with mixed service strategies* (Performance Evaluation, vol. 23, no. 2, pp. 107-120).

[208] Tijms, H.C., (1994). *Stochastic Models, an Algorithmic Approach* (John Wiley, New York).

[209] Titchmarsh, E.C., (1939). *The Theory of Functions, 2$^{\text{nd}}$ edition* (Oxford University Press, New York).

[210] Tsai, Z., Rubin, I., (1992). *Mean delay analysis of a message priority-based polling scheme* (Queueing Systems, vol. 11, pp. 223-240).

[211] Urlings, R.G.L.H., (2007). *Performance Evaluation of Monoether Factory at BASF AG* (Report, Eindhoven University of Technology).

[212] Vaughan, T.S., (2003). *The effect of correlated demand on the cyclical scheduling system* (International Journal of Production Research, vol. 41, no. 9, pp. 2091-2106).

[213] Vergin, R.C., Lee, T.N., (1978). *Scheduling rules for the multiple product single machine system with stochastic demand.* (Journal of Operational Research and Information Processing, vol. 16, no. 1, pp. 64-73).

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

[215] Vught, R., (2005). *Dakpannen op de Grond: Hoeveel & Hoe Minder* (Master's thesis, Eindhoven University of Technology, in Dutch).

[216] Vuuren, M. van, Adan, I.J.B.F., (2006). *Performance analysis of assembly systems* (Proceedings of the Markov Anniversary Meeting, pp. 89-100).

[217] Vuuren, M. van, Adan, I.J.B.F., Resing-Sassen, S.A., (2005). *Performance analysis of multi-server tandem queues with finite buffers and blocking* (OR Spectrum, vol. 27, no. 2-3, pp. 315-338).

[218] Wagner, M., Smits, S.R., (2004). *A local search algorithm for the optimization of the stochastic economic lot scheduling problem* (International Journal of Production Economics, vol. 90, pp. 391-402).

[219] Whitt, W., (1986). *Approximating a point process by a renewal process I: two basic methods* (Operations Research, vol. 30, pp. 125-147).

[220] Wolff, R.W., (1982). *Poisson arrivals see time averages* (Operations Research, vol. 30, no. 2, pp. 223-231).

[221] Zangwill, W.I., (1992). *The limits of Japanese production theory* (Interfaces, vol. 22, no. 5, pp. 14-25).

[222] Zangwill, W.I., (1994). *Response to comments on our work by Duenyas, by Gerchak and Zhang, and by McIntyre* (Interfaces, vol. 24, no. 5, pp. 90-94).

[223] Zheng, Y., Zipkin, P.H., (1990). *A queueing model to analyze the value of centralized inventory information* (Operations Research, vol. 38, no. 2, pp. 296-307).

[224] Zipkin, P.H., (1986). *Models for design and control of stochastic multi-item batch production systems* (Operations Research, vol. 34, no. 1, pp. 91-104).

[225] Zipkin, P.H., (1995). *Performance analysis of a multi-item production-inventory system under alternative policies* (Management Science, vol. 41, pp. 690-703).

[226] Zipkin, P.H., (2000). *Foundations of Inventory Management* (McGraw-Hill, London).

Summary

Polling, Production & Priorities

The present monograph focuses on the so-called *stochastic economic lot scheduling problem* (SELSP), which deals with the make-to-stock production of multiple standardized products on a single machine with limited capacity under *random demands*, *possibly random setup times* and *possibly random production times.* In the SELSP, a production policy is needed which describes for each possible state of the system whether to continue production of the current product, whether to switch to another product or whether to idle the machine. The objective of the present monograph is the development and the analysis of mathematical models that capture the behavior of the class of *fixed-sequence base-stock policies.* For given base-stock levels, it is shown that the analysis of a fixed-sequence base-stock policy is tantamount to the analysis of the queue length distribution in a classical queueing model, the so-called *polling system.*

The focus of the current research is mainly on the lot-sizing decision: what should the length of the production run be? Within the context of this lot-sizing decision the present monograph is, in particular, concerned with the evaluation and comparison of the traditional exhaustive and gated lot-sizing policies, on the one hand, and the more sophisticated quantity-limited lot-sizing policy, on the other hand. The latter offers the possibility to prioritize among the different products for improving total system performance through bounding the lengths of the production runs. Evaluation and optimization of these lot-sizing disciplines are achieved through state-of-the-art analysis of several polling systems. We study two research objectives as summarized below.

Research objective 1. *Development of a unifying exact framework for the analysis of the exhaustive and gated lot-sizing policies in terms of the average work-in-progress (WIP) levels under the assumption of Poisson demand processes.* □

In Chapter 3 an exact *Mean Value Analysis* (MVA) framework for the exhaustive and gated lot-sizing disciplines is presented, which computes the average WIP levels by exploiting direct mean value arguments. Within this framework the individual WIP levels can be efficiently obtained via the solution of a sparse set of linear equations, whereas for the total WIP level a closed-form expression is presented.

The MVA framework gives rise to explicit closed-form expressions, allowing for back-of-the-envelope calculations, for the individual WIP levels in the asymptotic regime of high utilization of capacity due to either customer demands or setup times. These expressions explicitly show the impact of all input parameters, yield insensitivity and monotonicity properties and unearth the (dis)similarities between the two sources of high utilization. In particular, it is shown that the exhaustive and gated lot-sizing disciplines display undesirable behavior if the utilization rate is high due to customer demand, which reveals itself, for example, in difficulties in the coordination between stages within the production process.

Motivated by the practical significance of the large setup times regime, we study this regime in more detail for a general class of branching-type lot-sizing policies by using more advanced techniques. The most remarkable result of this analysis is the fact that the stochastic system converges to its deterministic counterpart in the limit of increasing setup times implying that the exhaustive lot-sizing policy is optimal in terms of the WIP levels and that, thus, production runs should not be bounded in systems with extremely large setup times. For general settings, the latter conclusion does not always hold which we analytically show in the analysis of the second research objective.

Research objective 2. *Development of an efficient and accurate approximate tool for the analysis of the quantity-limited lot-sizing policy under the assumption of general demand processes.* □

In order to gain insights into the impact of bounding production runs and not to be diverted by other effects, Chapter 4 starts the analysis with a basic occurrence of the SELSP in an exact way. That is, we analyze a two-product system, in which a *high-priority* product is produced *exhaustively* and a low-priority product according to the *quantity-limited* service strategy. In this model, we observe significant cost reductions by application of the quantity-limited policy, compared to the standard exhaustive policies, indicating the potential of the quantity-limited service discipline as lot-sizing rule in production environments.

The results obtained in the two-product case provide us with theoretical evidence that the quantity-limited strategy may lead to considerable cost reductions compared to the widely used (standard) exhaustive policy. Therefore, in Chapter 4 we develop an efficient and accurate approximate decomposition approach for the evaluation of quantity-limited lot-sizing policies under the most general imaginable assumptions, i.e., general number of products each with their own quantity limit in an environment with generally distributed arrival, service time and setup time distributions. The accuracy of the approximation scheme is verified by means of an extensive simulation study.

The last part of Chapter 4 is devoted to a numerical simulation study assessing the quality of the quantity-limited lot-sizing policy as tool for prioritizing among products. It is shown that the quantity-limited lot-sizing policy outperforms the standard exhaustive policy leading to improvements in system performance for a variety of environments.

Finally, we would like to emphasize that the results of the present monograph are certainly not limited to the described production setting, but may be used in the design and optimization phase of many other fields of application such as communication, maintenance, manufacturing and transportation.

About the author

Erik Mathias Maria Winands was born in Valkenburg (The Netherlands) on February 10, 1980. He received his M.Sc. degree in Industrial and Applied Mathematics (cum laude) from Eindhoven University of Technology in 2003. Subsequently, he became a Ph.D. student at the same university (Department of Mathematics and Computer Science and Department of Technology Management) under supervision of Onno Boxma and Ton de Kok. Erik defends his thesis on September 4, 2007. As of September 2007, he will be working as risk validator at Rabobank.