Approximate analysis of queueing network models

approximate analysis of queueing network models

j.b.m. van doremalen

approximate analysis of queueing network models

approximate analysis of queueing network models

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE TECHNISCHE WETENSCHAPPEN AAN DE TECHNISCHE HOGESCHOOL EINDHOVEN, OP GEZAG VAN DE RECTOR MAGNIFICUS, PROF. DR. F. N. HOOGE, VOOR EEN COMMISSIE AANGEWEZEN DOOR HET COLLEGE VAN DEKANEN IN HET OPENBAAR TE VERDEDIGEN OP VRIJDAG 14 MAART 1986 TE 16.00 UUR.

DOOR

JOHANNES BERNARDUS MARIA VAN DOREMALEN

GEBOREN TE HEDEL

Druk: Dissertatiedrukkerij Wibro, Helmond.

Dit proefschrift is goedgekeurd
door de promotoren

Prof.dr. J. Wessels

en

Prof.dr. J. Wijngaard

CONTENTS

1. INTRODUCTION AND SUMMARY

1.1. Queueing network models.

Networks of queues are used in many areas to model and analyse real-life systems, as for instance in computer performance analysis, design of communication networks, production planning in manufacturing enterprises and planning of transportation systems. The use of such models is justified by a common characteristic of such systems: they can be viewed as a collection of interconnected resources providing service to a group of users. The resources have a finite capacity and, consequently, waiting lines of jobs may be formed.

Especially for the performance evaluation and the design of information processing systems queueing network models have proved to be a reliable and accurate tool for analysis, cf. Kleinrock [1975:1] and [1975:2], Ferrari [1978], Kobayashi [1978], Gelenbe and Mitrani [1980], Sauer and Chandy [1981], Lavenberg [1983] Lazowska, Zahorjan, Graham and Sevcik [1984] and Heidelberger and Lavenberg [1985].

Mathematical models provide an efficient tool for the evaluation of large and complex queueing network systems. The development of an adequate model assumes a thorough study of the essential features of the system and as such provides a clear and profound insight in the operation of the system. The mathematical analysis of the model may lead to the development of an efficient tool for the evaluation of the system for different parameter sets and as such yields the possibility of parametrization.

The analysis of mathematical models yields thus an attractive basis for the design, control and performance improvement of queueing network systems.

1.2. The analysis of queueing network models

Standard tools in the study of models of queueing network systems are simulation and mathematical analysis. In some instances "hybrid methods" have to be advocated.

It is typical for simulation that it may be used to analyse mathematical models at virtually every level of detail. Essentially, it provides a tool for the analysis of transient or time-dependent behaviour by the construction of sample paths of the underlying stochastic process. In practice, the use of simulation is restricted as the evaluation of large and complex models by simulation takes large amounts of computation time and demands vast storage requirements. Furthermore, the statistical interpretation of the numerical results is a difficult and quite often underestimated problem. In Bratley,Fox and Schrage [1983] an overview is given of the technical aspects of simulation. The simulation of queueing network models for instance has been treated in Markowitz [1983], Sauer and Chandy [1981] and Kobayashi[1978].

Typical for analytical techniques is that these may yield efficient algorithmic procedures for the computation of important system characteristics. They, essentially, provide a tool for the analysis of the equilibrium behaviour. The major problem is that only for a

relatively small class of queueing network models exact and efficient evaluation techniques are known.

Both simulation and analytical techniques have their restrictions. For large and complex queueing network models neither of the two techniques provides an efficacious tool. For such models the development of approximation methods seems to be a natural way-out.
In this monograph we concentrate on the development of analytical approximation methods. It is our aim to show the lines of argument which may lead to the development of intuitive, accurate and efficient approximation methods. Some more or less well-established problems from the area of queueing network analysis illustrate the general ideas.

1.3. Separable queueing network models

Stochastic models and particularly Markov processes are a widely used tool for the analysis of queueing network systems. Though it is, in principle, possible to study the transient behaviour of such models, in practical situations one concentrates on the analysis of the limiting behaviour. Under certain ergodicity conditions the equilibrium and limiting distribution of a Markov process may be obtained as the unique, strictly positive and normalized solution of a set of linear equations relating the equilibrium or limiting probabilities.
If a queueing network model apart from these ergodicity conditions satisfies so-called separability conditions, the solution of this set of equations attain an attractive product form. Such models are therefore called separable or product form queueing network models.
This line of research started in Jackson [1957] in which it has been proved that the equilibrium distribution of a particular type of network models has a product form. Extension of Jackson's results has shown that in a larger class of networks the set of equilibrium equations has a product form solution. Noteworthy papers in this line are Gordon and Newell [1967] and Baskett, Chandy, Muntz and Palacios [1975]. The latter paper has set a standard for the class of product form or separable queueing network models: the BCMP network models. Further extensions have been studied for example in Lam [1977], Kelly [1976] and [1979], Hordijk and Van Dijk [1981] and [1984] and Lazar and Robertazzi [1984]. It seems that the boundaries of the class of separable queueing networks have been reached and that it is quite unlikely that substantial extensions are to be found any more.

Apart from the fact that for separable queueing network models analytical expressions for the equilibrium probabilities have been obtained, it has been shown that important system characteristics may be evaluated in a relatively simple and efficient way.
Queueing network models can be divided in open, closed and mixed open and closed networks. Open networks are characterized by the fact that the customers arrive from outside the system, proceed through the network and eventually leave it. In closed networks a fixed number of customers proceed through the system and the customers neither enter

nor leave the system. In mixed open and closed networks both types of customers are present.
For separable open queueing networks the evaluation of steady-state system characteristics is straightforward as each queue may be evaluated in separation: the network is separable in a strict sense. For separable closed queueing networks, and consequently for mixed open and closed networks as well, the analysis is more complicated. The product form establishes the equilibrium distribution up to a normalization constant. The evaluation of this constant causes computational problems, as it involves a summation of the unnormalized equilibrium probabilities over the complete state space.
The two main procedures for solving this problem are known as the convolution algorithm and the mean value analysis.

The convolution algorithm, introduced in Buzen [1973], is a recursive algorithm for the computation of the normalization constant. It appears that the system characteristics can be expressed in the recursively computed values and thus may be evaluated efficiently. In Reiser and Kobayashi [1975] the procedure has been extended to a large class of separable queueing network models.

The mean value analysis forms a recursive algorithm which is based on a set of relations between important system characteristics as for instance expected residence times, expected numbers of customers at the resources and throughputs. It has been introduced in Reiser [1979] and extended in Reiser and Lavenberg [1980], Zahorjan and Wong [1981], Krzesinski, Teunissen and Kritzinger [1982] and Bruell, Balbo and Afshari [1984]. The relations can be shown to hold by algebraic manipulation on the equilibrium probabilities. However, the relations have an attractive intuitive interpretation as well, as these may be viewed as consequences of two important results from queueing theory: Little's formula and an arrival theorem.
Little's formula expresses the expected number of customers in a queueing system as the product of the arrival rate of customers at that system and the expected residence time of a single customer in the system, cf. Little [1961] and Stidham [1974].
The arrival theorem couples the equilibrium distribution of the queueing network model with the equilibrium distributions at departure and arrival moments of individual customers, cf. Lavenberg and Reiser [1980] and Sevcik and Mitrani [1981].

The implementation of the convolution and mean value analysis algorithms is discussed for instance in Bruell and Balbo [1980], Chandy and Sauer [1980], Reiser [1981] and Zahorjan and Wong [1981].
A considerable improvement of the standard convolution algorithm algorithm may be accomplished by the use of the tree convolution algorithm which has been introduced in Lam and Lien [1983]. For the mean value analysis algorithm an analogous improvement, the tree MVA algorithm, has been suggested in Tucci and Sauer [1985].
Very recently a new convolution-like algorithm has been suggested in Conway and Georganas [1985]. It is claimed that for some types of queueing network models the method is considerably faster than the classical algorithms.

1.4 Approximate analysis of queueing network models

Regrettably, many realistic models for practical problems do not have a product form, i.e. are not separable, whereas separable models tend to be very large and therefore cannot be evaluated using standard methods. For both types of problems the use of approximation methods seems a way-out.

In this monograph particular emphasis will be put on the use of separable queueing network models and the computational procedures which are associated with such models.

A wave of papers has appeared in this area of research. We may, roughly, discern three approaches in the development of approximation methods. These will be discussed in more detail, but beforehand it should be noted that the approaches cannot be viewed as strictly independent. Whereas the last two approaches are typical for the approximate analysis of large and complex queueing network models, the first approach provides a more fundamental and theoretical framework.

The first approach is based on the analysis of the set of equilibrium equations of the Markov process. The emphasis is on the use of the sparseness and structure of the transition matrix which describes the linear system. Decomposition and aggregation are standard techniques for obtaining approximate solutions for such highly structured linear systems. Decomposition methods are based on a splitting of the state space of the Markov process in subsets of states. The analysis of an adjusted model at the states of a given subset is followed by a composition step in which the relevant information for the overall model is obtained. Such methods have been extensively dealt with in Courtois [1977] in which concepts as "exactly" and "nearly-completely-decomposable" matrices have been introduced. The approximation methods have an interesting algebraic background which makes it possible to formulate bounds for the evaluated approximations. Furthermore, the algebraic results may be interpreted in terms of the studied stochastic models. They may be viewed as consequences of the different time scales at which distinct parts of the system are operating. This interpretation provides a tool for constructing a decomposition of the state space which is based on the relevant features of the system rather than on the explicit structure of the transition matrix. Examples are discussed in Courtois [1977], Hine, Mitrani and Tsur [1979] and Kuehn [1979].
Aggregation methods are based on the lumping together of subsets of the state space in single states and the formulation of an adjusted model on this newly defined smaller state space. The analysis of this smaller model is followed by a disaggregation step in which the relevant characteristics of the original model are evaluated. Again, algebraic arguments may be used to obtain information on the accuracy of such approximating schemes, cf. the discussion on exact and approximate lumpability in Schweitzer [1984] and the notes on exact aggregation in Vantilborgh [1978] and Schassberger [1984]. In practice, the aggregation of the state space has to be a natural consequence of the structure of the underlying stochastic model.

It should be observed that both decomposition and aggregation methods may be implemented as iterative schemes. By iterating between the decomposition (aggregation) step and the composition (disaggregation) step one hopes to obtain better approximations.

The second approach is based on the theory of separable queueing networks. Since efficient computational procedures for such networks are available it seems natural to use separable queueing network models to approximate general or very large separable queueing network models. The problem is to find an appropriate set of parameters for the approximating model.
Most of the approximation methods are based on a hierarchical model of the queueing network system. At a higher level a relatively simple separable queueing network model describes the interaction between the components of the system. At a lower level more detailed models describe the operation of specific components of the system.
Decomposition and aggregation techniques are natural tools to analyse hierarchical models. Later on we will take up the question of when and how to use decomposition and aggregation techniques. Here we would like to point out the close relationship between the use of decomposition and aggregation in the first two approaches.
Typical examples in this line of research are the parametric analysis, introduced in Chandy, Herzog and Woo [1975:1] with applications in Chandy, Herzog and Woo [1975:2] and Chandy and Sauer [1978], and the iterative method introduced in Marie [1979] and Marie and Stewart [1977] with an application in Bondi [1984].

In recent years a third approach has attracted a lot of attention. Since the development of the mean value analysis and the understanding of its intuitive interpretation, a wave of papers has appeared on the use of this procedure as a basis for approximation methods for large and complex queueing network models. Though the methods are basically heuristic in nature, they show some structural resemblance for different types of problems. Apart from experimental results little is known about the accuracy of the proposed methods. However, the results are quite satisfactory and so the mean value analysis approach has become a popular and widely appraised tool, De Souza e Silva, Muntz and Lavenberg [1984], Lavenberg and Sauer [1983] and Van Doremalen and Wessels [1983].
The mean value analysis algorithm is recursive and thus the approximation methods may be formulated in a recursive fashion. In many applications an iteration is implemented anyway.
There are two good arguments to do so. The first one is that it may be attractive to iterate in order to capture more precisely the interactions between the components of the system. The second one is that iteration may be used to bypass the recursiveness of the mean value analysis procedure which causes serious computational problems for larger models.
One of the main conclusions of our research is that for closed queueing network models the implementation of a strictly recursive approximation method has to be advocated. Strictly recursive methods can be very efficient and accurate evaluation tools for the approximate analysis of large and complex queueing network systems.

We have not ventured on the question of when to use a specific approach and of how to design an approximation method. The structure of the model and the type of information that one wants to extract from the model are two important guides. In Chapter 2 we review some ideas with respect to the development of approximation methods for large and complex queueing network models. The examples which are presented in the chapters 4, 5 and 6, provide a further insight in the application of these ideas.

1.5 Aim of the monograph

In the preceding sections we have given a rough sketch of the exact and approximate analysis of Markovian models of queueing network systems.
In this monograph we sketch ideas and lines of reasoning which provide a basis for the development of efficient and accurate approximation methods for large and complex queueing network models. The main emphasis is on the use of the recursive mean value analysis algorithm and its intuitive interpretation.

The first conclusion of the research is that the mean value analysis provides an attractive basis for the development of highly efficient, accurate and appealing intuitive approximation methods.

The second conclusion is that for the approximate analysis of closed queueing network models the use of strictly recursive methods has to be advocated. Whereas in the recent literature the emphasis is on iterative approximation methods, it is our opinion that strictly recursive methods for several reasons form an interesting alternative.
In the first place strictly recursive methods use explicitly the structure of the mean value analysis reasoning which is based on a strict recursion in the number of customers. This idea of studying the system with a given number of customers and trying to estimate what happens when an extra customer is added forms an attractive basis for the development of approximation methods.
In the second place such methods appear to yield very good results that can be compared with related iterative methods.
In the third place the implementation of a strict recursion in a computer program has the considerable advantage that a priori estimates can be made for the amount of computation time needed and the size of the storage facilities needed.

This monograph is organized as follows. The first part concentrates on the characterization of some general ideas that may be useful in the development of approximation methods for queueing network models.
Important observations for the analysis of queueing network systems are that the models tend to be highly structured and that interest is in aggregated rather than in detailed information. Decomposition and aggregation techniques seem, therefore, obvious tools for the approximate analysis of such models. These techniques have to be based on the structure of a queueing network system and an understanding of the specific performance measures to be evaluated. We shall discuss these global observations in more detail later

on.
For the approximate analysis of large queueing network models we have decided to use separable queueing network models and, more in particular, the mean value analysis procedure with its appealing intuitive interpretation. We shall discuss the use of these techniques in more detail and provide an introduction to the analysis of separable queueing network models and the mean value analysis procedure.

In the second part of the monograph the sketched ideas are elaborated for a set of queueing network models. The examples are typical in so far that they cover the essential problems that may arise in the analysis of large and complex queueing network models.

The first example considers the evaluation of the system characteristics of a closed multichain queueing network model. Computational complexity and storage requirements of the recursive mean value analysis procedure prohibit an exact evaluation for larger numbers of closed customer chains.
The second example considers a queueing network model with a special type of two-phase servers, where the first phase is a preparatory one. The problem is here the violation of the separability conditions.
The third example considers a closed queueing network with a priority schedule at one or more of the queues. Here we have to do with a combination of the problems: the computational complexity of the mean value analysis algorithm and the fact that the separability conditions are violated. problems.

1.6 Summary

This monograph deals with the development of approximation methods for the analysis of large and complex queueing network models. The first part, Chapters 2 and 3, deals with the introduction of some ideas and techniques which may be used as tools for the development of intuitive, efficient and accurate approximation methods. The second part, Chapters 4, 5 and 6, treats a number of worked out examples.

In Chapter 2 the development of approximation methods is discussed from a rather general point of view.
The first part of the chapter considers a description of a queueing network model in terms of a production structure, describing the operation of the resources, and a demand structure, describing the arrival processes, the routing behaviour and the service requirements of the customers. This structure in a natural way leads to the introduction of decomposition and aggregation techniques.
In the second part of the chapter the use of separable queueing network models and the associated mean value analysis procedure are discussed in more detail. The line of argument is augmented by some illustrating examples.

In Chapter 3 a class of mixed open and closed separable queueing network models with multiple customer chains and queue length dependent service rates is introduced. The

emphasis is on the mean value analysis procedure and its intuitive interpretation in terms of Little's formula and an arrival theorem. An implementation of the recursive algorithm is presented and remarks are made with respect to computational complexity and storage requirements.

Chapter 4 considers the numerical problem of evaluating the exact mean value analysis procedure for separable queueing network models with many closed customer chains. A classification of the approximation methods is presented and existing as well as new methods are described.

Chapter 5 treats the approximate analysis of a non-separable queueing network model with a special type of two-phase service units. The first part of the service is preparatory and can be performed in the preceding idle period for the first customer of the next busy period.

Chapter 6 deals with the approximate analysis of a queueing network model with preemptive resume and head-of-the-line priority queues. We present a new approximation method which is based on the mean value analysis procedure and a mean value analysis of M/G/1 priority queues.

2. THE APPROXIMATE ANALYSIS OF QUEUEING NETWORK MODELS

2.1 Introduction

This chapter is devoted to the introduction of some basic arguments which may lead to the development of efficient and accurate approximation methods for large and complex queueing network models.

In the first part of the chapter the use of decomposition and aggregation methods is reviewed.
The components of a queueing network model are split in a production structure, describing the operation of the resources or queues, and a demand structure, describing the arrival processes, routing behaviour and service demands of the customers or jobs. Though these two structures in general interfere in a more or less complicated way, the recognition of this basic distinction may lead to the design of efficient and accurate decomposition and aggregation methods.

The second part of the chapter deals with the use of separable queueing network models and the application of mean value arguments. Two approaches are discussed.
In the first approach non-separable queueing network models are approximated by separable ones and large separable models by smaller ones. The second approach consists of adjusting or extending the mean value analysis procedure using mean value arguments. For non-separable queueing network models this is done in such a way that the violation of the separability conditions is accounted for. For large separable queueing network models the computational complexity and storage requirements are diminished by considering smaller models.

The chapter is organized as follows. The distinction of a production and a demand structure and the use of decomposition and aggregation techniques are reviewed in Section 2.2. The use of separable queueing network models and the mean value analysis procedure is briefly discussed the Sections 2.4 and 2.5. The discussion is illustrated by some small examples which require the prerequisites provided in Section 2.3.

2.2. Decomposition and aggregation

In this section we discuss the use of decomposition and aggregation techniques in the development of approximation methods for large and complex queueing network models. The use of such techniques may be based on a description of a queueing network model in terms of a production structure and a demand structure.

The production structure describes the nature and operation of the resources. A resource is designed to execute a set of tasks in a specified way. Its operation may be expressed in terms of buffer capacities, service rates, service disciplines and availability.
The demand structure describes arrival processes, routing behaviour and service demands of the individual customers. Principally, each customer has its own set of tasks and it

needs a specific subset of resources to execute these tasks in a prescribed order. In practice, it is convenient to discern groups of customers which are alike. For each group or chain of customers one has to describe the structure of the set of tasks, the order in which the tasks have to be executed and the assignment of the resources that have to execute the tasks.

Note that a strict separation of the production and demand structure has been suggested. In many applications it is difficult or even impossible to make such a rigorous division; production and demand structure will interfere in a more or less complicated way. For the line of reasoning this is not a handicap, as the global distinction of a production and a demand structure in many instances leads to the development of efficient approximation methods.

Natural tools to go from a detailed description to a less detailed one are decomposition and aggregation. The basic idea of decomposition is to split a model with multiple resources into multiple models with single resources or to split a model with multiple groups of customers into multiple models with a single group of customers. The basic idea of aggregation is to lump groups of resources together into a single resource and several groups of customers into a single group of customers.

Observe that aggregation and decomposition may be used simultaneously, e.g. an aggregation of production structure and a decomposition of the demand structure can be combined in one method.
It should be noted that an aggregation step in general is followed by a disaggregation step and a decomposition step by a composition step. In approximation methods this leads to an iteration between aggregation and disaggregation step and between decomposition and composition step.

Let us next indicate the impact of performance and design questions on the choice of aggregation and decomposition methods. Performance and design questions can roughly be divided in production and demand oriented problems. Capacity planning and service discipline analysis are typical production oriented problems. Response time analysis is a typical example of a demand oriented problem.
It is obvious that many problems will fall in both categories. For example, if a new service discipline is considered, this will influence both the utilization of the resources and the response times of the customers. So production and demand oriented performance measures enter into the analysis. It will often be possible to isolate these problems and to analyse them separately.

In production oriented problems the demand structure is of secondary importance: it must be modelled in such a way that it reflects the influence of the demand processes on the distinct components of the queueing network model. An aggregation of the demand structure therefore seems a natural technique for approximating the influence of the demand structure on the production structure.
If the production oriented performance questions are directed towards the evaluation and

analysis of particular components or subsystem of components, then a decomposition of the production structure may be a useful tool. Note that this requires an adjustment of the demand structure as well.

In demand oriented problems the production structure has to be modelled in such a way that it reflects the influence of the production structure on the performance measures of the demand structure. An aggregation of the production structure is the obvious way to approximate this influence.
If the demand oriented problem is directed towards the analysis of specific groups of customers, the decomposition of the demand structure can be the next step in the approximate analysis.

In what follows the sketched arguments will be used explicitly and implicitly many times.

2.3. Two separable queueing network models

2.3.1. Introduction

In the next two sections the use of separable queueing network models and the mean value analysis procedure in the approximate analysis of queueing network models is reviewed. The line of reasoning is illustrated by a number of examples. In this section we introduce two relatively simple separable queueing network models that will be used in these examples.

2.3.2. A closed single chain queueing network model

The first queueing network model is a closed single chain queueing network with first-come first-served resources with a single service unit and queue length dependent service rates.
The production structure of the model comprises N resources, numbered $n = 1,2,...,N$. Each resource has an infinite buffer capacity. The single service unit at resource n, $n = 1,2,...,N$, has service rates $\mu_n(k)$ indicating the amount of work executed per unit time when k customers are present. The customers are serviced in order of arrival.
The demand structure of the model consists of K customers of a single type which proceed through the network in accordance with a Markov routing defined by an irreducible stochastic matrix P. The elements $p_{n,m}$, $n,m = 1,2,...,N$, describe the probability that a customer after its service has been completed at resource n joins the buffer at resource m. The service demands at resource n, $n = 1,2,...,N$, are stochastically independent and exponentially distributed with mean w_n.

Consider the following continuous-time Markov process. The state of the queueing network model is described by a vector $(k_1,k_2,...,k_N)$, where k_n, $n = 1,2,...,N$, denotes the number of customers at resource n. The state space S is the set of all states and is given

as,

$$S = \{ (k_1, \ldots, k_N) \mid k_n \in \{0,1,\ldots,K\}, n=1,2,\ldots,N, \text{ and } \sum_{n=1}^{N} k_n = K \}. \tag{2.3.1}$$

To be able to formulate analytic expressions for the equilibrium or limiting probabilities the following auxiliary quantities are introduced.
The visiting ratios f_n, $n=1,2,\ldots,N$, are the stationary distribution of the Markov chain defined by the routing matrix P, i.e the f_n's are the unique positive solution of a set of linear equations,

$$f_n = \sum_{m=1}^{N} f_m p_{m,n}, \; n=1,2,\ldots,N, \tag{2.3.2}$$

and a normalization equation

$$\sum_{n=1}^{N} f_n = 1. \tag{2.3.3}$$

The value f_n is called a visiting ratio as it may be interpreted as the stationary probability that the visit of a customer is to resource n.

In accordance with Theorem 3.1 the equilibrium probabilities $p(\mathbf{k})$, $\mathbf{k}=(k_1, \ldots, k_N) \in S$, satisfy,

$$p(\mathbf{k}) = \frac{1}{G} \prod_{n=1}^{N} F_n(k_n), \tag{2.3.4}$$

where, for $k=0,1,\ldots,K$,

$$F_n(k) = \prod_{i=1}^{k} \frac{w_n f_n}{\mu_n(i)} \tag{2.3.5}$$

and where

$$G = \sum_{\mathbf{k} \in S} \prod_{n=1}^{N} F_n(k_n). \tag{2.3.6}$$

Usually, one is not interested in these limiting probabilities but in steady-state characteristics, for example expected residence times, throughputs and expected numbers of customers. The characteristics for the equilibrium behaviour are informally introduced as follows

$S_n(K)$ expected residence time at resource n,

$\Lambda_n(K)$ expected number of arrivals per unit time at resource n, i.e. the throughput at resource n, and

$p_n(k, K)$ the probability that k customers are at resource n,

where K emphasizes the dependence of these behavioural characteristics on the population. In Chapter 3 these steady state characteristics are discussed in more detail and it is

shown that these may be evaluated in a relatively efficient way.

2.3.3. A closed multichain queueing network model

The second queueing network model is a closed multichain queueing network with first-come first-served resources with a single service unit and a fixed service rate.

The production structure consists of N resources. Each resource has an infinite buffer capacity and a single service unit that services the customers in order of arrival. The single service unit at resource n, $n=1,2,...,N$, has a fixed service rate which is normalized to unity.

The demand structure comprises R groups of customers or chains, numbered $r=1,2,...,R$. The K_r customers of chain r visit the resources of a subset $Q(r)\subset\{1,2,...,N\}$. The routing is defined by an irreducible and stochastic matrix P_r which is defined on $Q(r)\times Q(r)$. The elements $P_{r,n,m}$ denote the probability that a customer of chain r after a visit to resource n brings a visit to resource m. The service demands at resource n, $n=1,2,...,N$, are stochastically independent and exponentially distributed with mean w_n. This mean is independent of the chain number.

The visiting ratios $f_{n,r}$, are for each chain r, $r=1,2,...,R$, defined as the unique positive solution of

$$f_{n,r} = \sum_{m\in Q(r)} f_{m,r}\, p^r_{m,n} \quad , n\in Q(r) \tag{2.3.7}$$

and a normalization equation

$$\sum_{m\in Q(r)} f_{m,r} = 1 \, . \tag{2.3.8}$$

The following steady-state characteristics may be evaluated from a recursive scheme which is known as the mean value analysis algorithm. The characteristics are to be viewed as limiting quantities either for the number of visits going to infinity or for the time going to infinity. For $n=1,...,N$ and $r=1,...,R$ we introduce,

$S_{n,r}(\mathbf{K})$ expected residence time of chain r customers at resource n,

$\Lambda_{n,r}(\mathbf{K})$ throughput of chain r customers at resource n, and

$L_{n,r}(\mathbf{K})$ expected number of chain r customers at resource n.

where $\mathbf{K}=(K_1,...,K_N)$ denotes the dependence on the population vector.

In Section 3.3 an arrival theorem is derived that states that the limiting distribution of the state of the system upon an jump moment of a customer of a specific chain equals the limiting distribution as if one customer of this chain has been removed. In combination with Little's formula the arrival theorem offers an appealing intuitive basis for the

development of a recursive scheme for the evaluation of the performance measures. We sketch this intuitive derivation and refer to Chapter 3 for details.

Consider a customer of chain r, $r=1,2,...,R$, arriving at a resource $n \in Q(r)$. As a consequence of the arrival theorem, the expected number of customers of chain l that it will see in front of it, equals $L_{n,l}(\mathbf{K}-\mathbf{e}_r)$ where $\mathbf{e}_r$ is the r-th unit vector. Each of these customers is to be served before the arriving customer and, consequently, its expected residence time equals

$$S_{n,r}(\mathbf{K}) = \sum_{l=1}^{R} L_{n,l}(\mathbf{K}-\mathbf{e}_r)w_n + w_n \ . \tag{2.3.9}$$

The expected number of visits that customers of chain r, $r=1,...,R$, bring to resource m, $m \in Q(r)$, between two departures from a fixed resource n, $n \in Q(r)$ equals $f_{m,r}/f_{n,r}$. As a consequence the expected time between two departures from resource n of a chain-r customer equals

$$\sum_{m \in Q(r)} \frac{f_{m,r}}{f_{n,r}} S_{m,r}(\mathbf{K}) \ . \tag{2.3.10}$$

For the throughput of customers of chain r at resource n we thus obtain,

$$\Lambda_{n,r}(\mathbf{K}) = \frac{f_{n,r}K_r}{\sum_{m \in Q(r)} f_{m,r}S_{m,r}(\mathbf{K})} \ . \tag{2.3.11}$$

Finally, applying Little's formula to the single resource n and the single chain r yields for the expected numbers of customers

$$L_{n,r}(\mathbf{K}) = \Lambda_{n,r}(\mathbf{K})S_{n,r}(\mathbf{K}) \ . \tag{2.3.12}$$

Starting with $L_{n,r}(\mathbf{0})=0$, for $n=1,2,...,N$ and $r=1,2,...,R$, the Relations (2.3.9), (2.3.11) and (2.3.12) constitute a recursive scheme for the computation of the expected residence times, the throughputs and the expected numbers of customers.

2.4. Separable queueing network models

2.4.1. Introduction

If a queueing network model satisfies, apart from certain ergodicity conditions, so-called separability conditions, the equilibrium equations for the equilibrium probabilities of a detailed state description attain a product-form solution. In Section 1.3 it has been noted that efficient algorithms have been developed for the evaluation of some important steady state system characteristics.

Regrettably, in most realistic models of queueing network systems the separability conditions are not satisfied and, consequently, the corresponding efficient evaluation methods cannot be applied directly. On the other hand separable models can be very large and are therefore inaccessible for the standard evaluation methods.

The approximation of non-separable by separable models and of very large separable

models by simpler separable queueing network models is a standard way-out. The kernel of such methods is the construction of a separable queueing network model for which the characteristics are an accurate approximation of the corresponding characteristics in the original model.
As indicated in Section 1.4 these methods, in general, find their validation in the mathematical analysis of exact and approximate decomposition and aggregation techniques for large and structured sets of linear equations. In practice, the choice of a decomposition or aggregation has to be made after a study of the structure of the model and the explicit performance and design questions that have to be answered.
The next two subsections are devoted to an elaboration of these global ideas in two examples. In Subsection 2.4.2 an iterative aggregation/disaggregation technique based on an aggregation of the production structure is discussed for a non-separable queueing network model with non-exponential service demand distributions at resources with a first-come first-served service discipline. In Subsection 2.4.3 an iterative decomposition/composition method is discussed for a separable queueing network model with many closed customer chains.

2.4.2. Aggregation of the production structure

One of the ways to construct an approximating separable queueing network model is an aggregation of the complex or large production structure. In this subsection we discuss an iterative aggregation-disaggregation method for a queueing network model with a complex production structure and a relatively simple demand structure. The method is closely related to the parametric analysis method, cf. Chandy, Herzog and Woo [1975:1] and [1975:2], and the method introduced in Marie [1979] and Marie and Stewart [1977].

Consider a queueing network model with N resources. At the resources, the customer are serviced in order of arrival by a single service unit with a fixed service rate that is normalized to unity. The network is closed and the K customers form a single chain. The routing is defined by an irreducible stochastic matrix P. The service demands at resource n, $n = 1,...,N$, are stochastically independent and distributed in accordance with a distribution function G_n with mean w_n. The resulting model is non-separable. The separability conditions are violated by the non-exponentiality of the service demand distributions at resources with a first-come first-served discipline, cf. the discussion in Section 3.2.

The model is approximated by a separable model comprising N resources with a first-come first-served discipline and a single service unit with queue length dependent service rates $\mu_n(k)$, for $k = 1,...,K$ and $n = 1,...,N$. The service demands are assumed to be independent and exponentially distributed with unit mean. Note that this model has been sketched in Subsection 2.3.2. It is separable and, so, the performance measures may be computed efficiently.

The iterative aggregation-disaggregation method now proceeds as follows.

Assume that an initial set of service rates $\mu_n(k)$ has been given. A plausible first guess might be $\mu_n(k)=w_n^{-1}$. The corresponding separable queueing network model may be evaluated to obtain the relevant characteristics. These characteristics are approximations for the corresponding characteristics in the original non-separable model.
This completes the aggregation step. The relevant information for the disaggregation step is provided by the probabilities $p_n(k,K)$.

The disaggregation step starts with the evaluation of sets of queue length dependent instream rates $\lambda_n(k)$, $k=0,\ldots,K$, at each resource n, $n=1,\ldots,N$.
These instream rates are to be used as the input parameters for a procedure to evaluate the performance measures of a finite capacity M/G/1-queue with state-dependent instream rates. The outcomes of these procedures are used to construct a new set of queue length dependent service rates $\mu_n(k)$.

The instream rates $\lambda_n(k)$ are obtained from the analysis of the separable model. Note that $\lambda_n(k)$ describes the expected number of arrivals at resource n per unit time when k customers are present.
A fraction $p_n(k,K)$ of the time there are k customers present at resource n. So, the expected number of arrivals per unit time while k customers are present equals $\lambda_n(k)p_n(k,K)$. This number has to equal the expected number of service completions per unit time leaving k customers at resource n, i.e for $k=0,\ldots,K-1$,

$$\lambda_n(k)p_n(k,K) = \mu_n(k+1)p_n(k+1,K) \tag{2.4.1}$$

and, obviously, $\lambda_n(K)=0$.

A crucial point is the efficient evaluation of the M/G/1-queue with state dependent arrival rates. For suggestions on the evaluation of such queueing systems we refer to Marie [1980], Tijms and Van Hoorn [1981] and Van Hoorn [1983].
The evaluation of the new set of queue length dependent service rates varies with the structure of the problem.
Iterating between the analysis of the separable model and the analysis of the M/G/1-queues the characteristics of the original model are approximated.

In Chapter 5 we present an application of the method when analyzing a queueing network model with two-phase servers.

2.4.3. Decomposition of the demand structure

The second example uses a decomposition of the demand structure. The problems are caused by the high computational complexity and the large storage requirements of the exact evaluation method.

Consider the closed multichain queueing network model introduced in Subsection 2.3.3. The recursive scheme is given by the Relations (2.3.9), (2.3.11) and (2.3.12). It runs through all the vectors in the range of (0,...,0) up till $(K_1,\ldots,K_R)$. So, the number of

recursion steps equals,

$$\prod_{r=1}^{R}(K_r+1) \tag{2.4.2}$$

and a large part of the information has to be stored. It is evident that for larger values of $R, K_1, \ldots, K_R$ the computational complexity and storage requirements will prohibit an exact evaluation.

This problem has attracted a lot of attention in the recent literature and in Chapter 4 it is analysed in detail. As an example we present an iterative decomposition/composition method which is based on a decomposition of the demand structure.

Consider a set of R separable single chain queueing network models. The rth network is associated with chain r, $r=1,2,\ldots,R$, in the original model. It is characterized by the set $Q(r)$ of resources and the K_r customers of chain r with routing matrix P_r. The routing behaviour is left intact, but the expected service demands at the resources are adjusted to account for the influence of the remaining chains. For the chain-r customers we introduce at each resource n, $n \in Q(r)$ an adjusted expected service demand $w_{n,r}$.

The method now proceeds as follows. Assume that an initial set of adjusted expected service demands $w_{n,r}$, $n=1,\ldots,N$ and $r=1,\ldots,R$, has been given. For each chain r, $r=1,2,\ldots,R$, the performance measures are evaluated from a one-dimensional recursive scheme, cf. Subsection 2.3.3. For $k=1,\ldots,K_r$ evaluate at all $n \in Q(r)$

$$S_{n,r}(k) = L_{n,r}(k-1)w_{n,r} + w_{n,r}\ , \tag{2.4.3}$$

$$\Lambda_{n,r}(k) = \frac{f_{n,r}k}{\sum_{m \in Q(r)} f_{m,r}S_{m,r}(k)}\ , \tag{2.4.4}$$

$$L_{n,r}(k) = \Lambda_{n,r}(k)S_{n,r}(k)\ . \tag{2.4.5}$$

Observe that the number of recursion steps of this scheme equals,

$$\sum_{r=1}^{R}(K_r+1)\ . \tag{2.4.6}$$

The problem is now to find new and hopefully better values $w_{n,r}$ in such a way that the performance measures of the proposed scheme approximate the corresponding measures in the original model.

For this step in the approximation method we perform a composition step. It seems reasonable to consider an adjustment of the effective expected service demand w_n accounting for the influence of the remaining chains. A plausible guess is to set $w_{n,r}$ to

$$w_{n,r} = \frac{w_n}{1-\sum_{l \neq r}\Lambda_{n,l}(K_l)w_n}\ . \tag{2.4.7}$$

Note that in this way the resource n appears to the customers of chain r as a resource which is slowed down by a factor corresponding with the fraction of time it is serving

customers

2.5. Mean Value Analysis

2.5.1. Introduction

In recent years the use of mean value analysis arguments has become a widely advocated and appraised basis for the development of approximation methods. Since the presentation of the mean value analysis procedure and its attractive interpretation in terms of an arrival theorem and Little's formula a wave of papers has appeared on the subject, see e.g. De Souza e Silva, Muntz and Lavenberg [1984] and Lavenberg and Sauer [1983] for an overview and discussion. The reason for this success is threefold.
In the first place, the appealing interpretation in terms of an arrival theorem and Little's formula yields an attractive basis for the development of adjusted and extended mean value analysis procedures. Such procedures may be used for the evaluation of approximations for performance characteristics in non-separable and large separable queueing network models.
In the second place, these adjusted procedures can be implemented with relative ease in existing mean value analysis algorithms.
In the third place the approximation methods have proved their value in the analysis of queueuing network systems, especially in the field of computer system and communication network analysis.

As we have seen in Subsection 2.3.3 the mean value analysis procedure comprises relations between important steady-state system characteristics, viz. relations for the expected residence times (2.3.9), for the throughputs (2.3.11) and for the expected numbers of customers (2.3.12). The latter two relations may be viewed as consequences of Little's formula and will hold in most non-separable networks as well. The expected residence time relation is special as it is based on the arrival theorem which is typical for separable queueing network models.
Thus, the crucial steps in the design of an approximating mean value analysis procedure are the formulation of an adjusted or extended relation for the expected residence times and the introduction of an approximating arrival theorem. The methods will therefore show for different types of problems some structural resemblance, but are typically heuristic in nature. Apart from experimental results little is known about the accuracy of the proposed methods. In several practical situations the results have appeared to be quite satisfactory.

In the next two subsections we give examples illustrating the main lines which may be followed in the design of mean value analysis based approximation methods. The first example considers a trictly recursive mean value analysis extension, whereas the second example presents an iterative method.

2.5.2. Non-exponential service demand distributions

Consider the queueing network model introduced in Subsection 2.3.3. However, now the service demands of customers of chain r, $r=1,2,...,R$, at resource n, $n=1,2,...,N$, have a distribution function $G_{n,r}$ with mean $w_{n,r}$ and variance $\sigma_{n,r}^2$. This network is non-separable and the arrival theorem cannot be invoked to construct an expected residence time relation.

A first approximation for the evaluation of the expected residence times in this non-separable queueing network is the following one. We use an adjustment of (2.3.9)

$$S_{n,r}(\mathbf{K}) = \sum_{l=1}^{R} L_{n,l}(\mathbf{K}-\mathbf{e}_r)w_{n,l}+w_{n,r} \ , \tag{2.5.1}$$

as if the arrival theorem holds (which certainly is not the case) and, consequently, a customer of chain r would upon its arrival at resource n see an expected number of $L_{n,l}(\mathbf{K}-\mathbf{e}_r)$ customers of chain l. The mean service demand of a chain-l customer is $w_{n,l}$. Summing these mean service demands and adding the mean service demand of the customer itself yields Relation (2.5.1).

The approximation may be improved. If a customer of chain l is being served upon the arrival moment of a customer of chain r, its expected residual service demand, in general, does not equal $w_{n,l}$. It seems a better guess to apply the mean residual life time formula at an arbitrary time instant, cf. Kleinrock [1975:1]. So, if a customer of chain l is in service, we use

$$\frac{\sigma_{n,l}^2 + w_{n,l}^2}{2w_{n,l}} \tag{2.5.2}$$

to approximate the expected residual service demand at the arrival moment of a customer of chain r.
Note that $\Lambda_{n,l}(\mathbf{K}-\mathbf{e}_r)w_{n,l}$ is an obvious approximation for the probability that a customer of chain r upon its arrival at resource n finds a customer of chain l in service. This yields the following improvement of Relation (2.5.1),

$$\begin{aligned} S_{n,r}(\mathbf{K}) = & \sum_{l=1}^{R} (L_{n,l}(\mathbf{K}-\mathbf{e}_r) - \Lambda_{n,l}(\mathbf{K}-\mathbf{e}_r)w_{n,l})w_{n,l} \\ & + \sum_{l=1}^{R} \Lambda_{n,l}(\mathbf{K}-\mathbf{e}_r)w_{n,l}\frac{\sigma_{n,l}^2+w_{n,l}^2}{2w_{n,l}} + w_{n,r} \ . \end{aligned} \tag{2.5.3}$$

The similarity between this formula and the Pollaczek-Khintchine formula for M/G/1-queues should be observed, cf. Oliver [1964] and Kleinrock [1975:1] for similar reasonings to obtain the expected residence times in M/G/1-queues.

Note that (2.5.3) in combination with (2.3.11) and (2.3.12) forms a strictly recursive scheme for the evaluation of approximations for the performance measures. It is easily verified that, in a straightforward way, the scheme can be implemented in any existing mean value analysis algorithm.

In Chapter 6 the proposed technique shall be applied in the more complex setting of a queueing network model with a preemptive resume priority schedule at some of the resources.

2.5.3 Closed multichain queueing network models

As we have seen in subsection 2.4.3 the main problem in the evaluation of the performance measures in large separable queueing network models with many closed customer chains are the high computational complexity and the large storage requirements of the mean value analysis algorithm.

In this subsection an alternative approximation method is highlighted which is based on an adjustment of the mean value analysis procedure. For more details we refer to Chapter 4. The simple and iterative method has been examined in Chow [1983]. The idea is to remove the recursion from the mean value analysis procedure by assuming that customers arriving at a resource see the system as if in equilibrium.

The resulting approximate scheme has the following form,

$$S_{n,r}(\mathbf{K}) = \left(\sum_{l=1}^{R} L_{n,l}(\mathbf{K}) + 1\right) w_n \ , \tag{2.5.4}$$

$$\Lambda_{n,r}(\mathbf{K}) = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(\mathbf{K})} \ , \tag{2.5.5}$$

$$L_{n,r}(\mathbf{K}) = \Lambda_{n,r}(\mathbf{K}) S_{n,r}(\mathbf{K}) \ . \tag{2.5.6}$$

This yields a set of non-linear equations for the approximate performance measures at the population vector K.

Using Brouwer's fixed point theorem one may show the existence of a solution of this set of equations. A standard way to obtain such a fixed point is by successive approximations. In Chow [1983] it is shown that the resulting iterative scheme converges to a unique fixed point.
Brouwer's fixed point theorem and the successive approximation method are standard tools in the analysis of iterative extensions of the mean value analysis procedure. Regrettably, little is known with respect to convergence of the iteration schemes and the uniqueness of the fixed points, confer De Souza a Silva, Muntz and Lavenberg [1984] and Lavenberg and Sauer [1983]. In Chapter 4 this fixed point problem is revisited in the context of approximation methods for separable queueing networks with many closed customer chains. In Chapter 5 we discuss the problem in a different setting, when studying a queueing network model with a special type of two-phase servers.

3. MIXED OPEN AND CLOSED MULTICHAIN QUEUEING NETWORK MODELS

3.1. Introduction

In Chapter 1 we have briefly reviewed the relevant literature on separable queueing network models. It has been noted that two approaches can be distinguished in the development of efficient computational procedures for the evaluation of system characteristics: the convolution method and the mean value analysis. In this chapter the Mean Value Analysis (MVA) approach is presented.

The MVA approach is based on an arrival theorem and Little's formula and allows for an intuitive derivation of a set of relations between system characteristics as expected residence times, throughputs, expected numbers of customers at the resources. These relations are the basis of the recursive Mean Value Analysis (MVA) algorithm.
Regrettably, the MVA approach does not prove the correctness of the relations. A technical proof based on algebraic manipulation of the detailed steady state probabilities must back up the approach. A detailed presentation of the proof falls outside the scope of this monograph and we content ourselves therefore with a number of references.
That the MVA approach is nevertheless presented, is suggested by the discussions in Section 1.4 and 2.5. It provides an attractive tool for the development of approximation methods.

The chapter is organized as follows.

In Section 3.2 a mixed open and closed multichain queueing network model is introduced as an irreducible, aperiodic and time-homogeneous Markov process on a finite or denumerable state space. Ergodicity and separability conditions are formulated that guarantee the existence and the product form of the limiting distribution of the Markov process.

In Section 3.3 we present the MVA approach. The first part of the section is devoted to an introduction of the steady-state system characteristics and the formulation of the relevant relations between these characteristics. Then we formulate and prove an arrival theorem and discuss the use of Little's formula.
The second part is devoted to the MVA approach. First, the approach is demonstrated in two special cases, an open and a closed queueing network model each with one customer chain. Then, then mixed multichain queueing network model with queue length dependent service rates is treated. It is also shown that the reasoning simplifies for models with fixed service rate resources.

In Section 3.4 an implementation of the MVA algorithm is described and remarks on the computational complexity and storage requirements of the algorithm are made.

3.2. A class of separable queueing network models

3.2.1. Introduction

In this section a class of mixed open and closed multichain queueing network models is introduced. The existence and structure of the limiting distribution of an associated continuous-time Markov process is discussed.

In Subsection 3.2.2 the queueing network model is introduced. The production structure comprises a set of resources. The versatile description of the resources allows for the introduction of well-known service disciplines as first-come first served, processor sharing and infinite server. The demand structure comprises open as well as closed customer chains.

The queueing network model generates an irreducible, aperiodic and time-homogeneous Markov process with a finite or denumerable state space: the buffer occupation process. In Subsection 3.2.3 we formulate an ergodicity condition guaranteeing the existence of the limiting probabilities and a separability condition guaranteeing the solution of the equilibrium equations to have a product form solution if any.
Queueing network models with this type of analytical solutions for the set of equilibrium equation are called product-form or separable queueing networks models.

3.2.2. A mixed open and closed multichain queueing network model

We introduce the production and demand structure of a mixed open and closed queueing network model.

The production structure comprises N resources, numbered $n = 1,2,...,N$. The set of resource is also denoted by $Q = \{1,...,N\}$. The buffers for these resources have an infinite capacity, so there will be no blocking due to full buffers. Furthermore, it is assumed that the resources are not subject to interrupts or breakdowns, i.e. the availability is 100 %.

The service rate at resource n, $n \in Q$, is given by a service rate function μ_n, where $\mu_n(k)$ indicates, for $k = 1,2,...$, the service rate if k customers are present at resource n.

As in Kelly [1976] and [1979] the service discipline at a resource is described by an admittance and a dedication function.

The admittance function γ_n indicates at which buffer place an arriving customer is stored. If $K-1$ customers are present at resource n, an arriving customer is stored at buffer-place k, $k = 1,2,...,K$, with probability $\gamma_n(k,K)$. The customers in the places $k,...,K-1$ are shifted to the places $k+1,...,K$.
The dedication function ϕ_n distributes the service capacity over the customers present at resource n. If K customers are present, a fraction $\phi_n(k,K)$ of the service rate $\mu_n(K)$ is dedicated to the customer at buffer place k, $k = 1,2,...,K$. At the moment the job of a

customer at buffer place k has been completed, the customer leaves the buffer and the customers at the places $k+1,\ldots,K$ are shifted to the places $k,\ldots,K-1$.
For every $K=1,2,\ldots$ admittance and dedication functions form proper distribution functions, i.e.

$$\sum_{k=1}^{K} \phi_n(k,K) = \sum_{k=1}^{K} \gamma_n(k,K) = 1 .$$

The modelling in terms of service capacity, admittance and dedication functions includes several well-known service disciplines. We discuss the first-come first-served, processor sharing and infinite server service disciplines. It does not allow for all types of service disciplines, as for instance priority schedules or blocking phenomena.

At a first-come first-served resource the customers are served in order of arrival. Let c be the number of service units and μ the constant service rate of each of these units. The service discipline is then modelled by the following service-rate, admittance and dedication functions

$$\mu(k) = \begin{cases} k\mu & , 1 \leqslant k \leqslant c \\ c\mu & , c \leqslant k \end{cases}$$

$$\gamma(k,K) = \begin{cases} 1 & , k=K \\ 0 & , \text{otherwise} \end{cases}$$

$$\phi(k,K) = \begin{cases} \frac{1}{k} & , 1 \leqslant k \leqslant K \leqslant c \\ \frac{1}{c} & , 1 \leqslant k \leqslant c \leqslant K \\ 0 & , \text{otherwise} \end{cases}$$

The processor-sharing service discipline is characterized by the fact that the total service rate is evenly distributed over the customers present. Thus, it has no influence on the operation of the resource at which buffer place an arriving customer is stored. We may therefore put, for given K and $k=1,\ldots,K$,

$$\phi(k,K) = \gamma(k,K) = \frac{1}{K} , k=1,2,\ldots,K .$$

The infinite server discipline may be viewed as a special case of the processor sharing discipline. An infinite number of service units each with a constant service-rate of μ serve the customers. The service rate function is given by $\mu(k)=k\mu$, $k=1,2,\ldots$. The admittance and dedication function are as in the processor sharing case.

The demand structure comprises open and closed customer chains. We restrict ourselves to a relatively simple modelling of arrival processes, routing behaviour and service demands. Some extensions are discussed at the end of Subsection 3.2.3.

We discern L customer chains, numbered $r = 1,...,L$. The chains $r = 1,...,R$ are closed and the chains $i = R+1,...,L$ are open. We use $C = \{1,...,R\}$ and $O = \{R+1,...,L\}$ to denote the sets of closed and open customer chains respectively.

Let us first discuss the modelling of a closed customer chain r, $r \in C$, containing K_r customers.
The routing is described by a stochastic matrix P_r, which is defined on $Q \times Q$. The element $P_{r,m,n}$, $n,m \in Q$, is the probability that a customer after a visit to resource m joins the buffer at resource n. The matrix P_r may be viewed as the transition matrix of an embedded Markov chain on the finite set Q. It is assumed that this Markov chain has one recurrent class. The row vector f_r of stationary probabilities of this Markov chain is the unique solution of the set of equilibrium equations and a normalization, viz.

$$f_r = f_r P_r \text{ and } f_r e = 1 , \tag{3.2.1}$$

where e is a column vector of ones.
The component $f_{n,r}$, $n \in Q$, of f_r is called a visiting ratio, since it may be interpreted as the stationary probability that the visit of a customer is to resource n.
It should also be observed that $f_{m,r} / f_{n,r}$ is the expected number of visits to resource m between two successive visits to a particular resource n. This interpretation will be of importance in the discussion of the MVA approach later on.

The service demands at resource n are stochastically independent and exponentially distributed random variables with mean $w_{n,r}$.

Let us next discuss the modelling of an open chain i, $i \in O$. The customers arrive at the system in accordance with a Poisson process with rate λ_i and join the buffer of resource $n \in Q$ with probability $p_{n,i}$. The probabilities $p_{n,i}$ form a row-vector $\boldsymbol{p}_i$.
The routing is described by a substochastic matrix P_i with elements $P_{i,m,n}$, $n,m \in Q$, being the probabilities that customers after a visit to resource m join the buffer at resource n. Note that with probability

$$1 - \sum_{n=1}^{N} P_{i,m,n} \tag{3.2.2}$$

a customer leaves the system after a visit to resource m.
The substochastic matrix P_i may be viewed as the transition matrix of an embedded Markov chain on the set Q. It is assumed that all the states of this chain are transient. This implies that all customers entering the system will eventually leave. We now may define the row-vector Λ_i as the unique solution of the linear system

$$\Lambda_i = \lambda_i \boldsymbol{p}_i + \Lambda_i P_i . \tag{3.2.3}$$

The solution of this linear system has an important interpretation. Under certain conditions, which for example guarantee the non-degeneracy of the limiting behaviour of the queueing processes, the component $\Lambda_{n,i}$, $n = 1,...,N$, of Λ_i may be interpreted as the limiting rate of the arrival or departure process of chain i customers at resource n. Therefore,

this quantity shall be referred to as the throughput of chain i customers at resource n.

The service demands at resource n are stochastically independent and exponentially distributed random variables with mean $w_{n,i}$.

This completes the formulation of the model.

3.2.3. Buffer occupation process and product form solution

Let $\{X(t), t \geqslant 0\}$ be a Markov process describing in a detailed way the stochastic behaviour of the buffer occupations. For every $t \geqslant 0$ $X(t)$ takes values in a state space S. The elements s of S are represented as row-vectors $s = (s_1,..,s_N)$, where $s_n, n \in Q$, describes the buffer occupation at resource n as

$$s_n = (k_n, r_n(1), \ldots, r_n(k_n)) . \tag{3.2.4}$$

Here, k_n denotes the number of customers at resource n and $r_n(k)$ the chain to which the customer at buffer place k belongs. If the buffer is empty, we set $s = (0)$.
The state space S comprises all possible states s. We write $S(K)$ instead of S to emphasize the dependence of S on the population vector K defined as $K = (K_1, \ldots, K_R)$.

The definitions of the production and demand structure guarantee the process $\{X(t), t \geqslant 0\}$ to be an irreducible, aperiodic and time-homogeneous Markov process on a finite or denumerable state space. For such processes one may analyse the behaviour for large t. If a limiting distribution exists, the limiting probabilities $p(s, K)$ are the unique, normalized and strictly positive solution of a set of global balance or equilibrium equations cf. Chung [1960], Cinlar [1975], Ross [1982] and Heyman and Sobel [1982]. These limiting probabilities are, for all $s \in S(K)$, defined as

$$p(s, K) = \lim_{t \to \infty} Pr\{X(t) = s\} . \tag{3.2.5}$$

In the sequel of this section we shall treat the existence and the product form of the solution of the set of global balance equations for the buffer occupation process $\{X(t), t \geqslant 0\}$. It falls outside the scope of this monograph to discuss the proofs at length. For details on the analysis and the methods of proof we refer to Baskett, Chandy, Muntz and Palacios [1975], Reiser and Kobayashi [1975], Kelly [1976] and [1979], Chandy, Towsley and Howard [1977] and Hordijk and Van Dijk [1981].

We start with the formulation of ergodicity and separability conditions which guarantee the existence and product form of the limiting distribution, respectively.

The ergodicity condition.
The limiting distribution exists if at all resources $n = 1,...,N$

$$\sum_{k=1}^{\infty} \prod_{l=1}^{k} \frac{\sum_{i \in O} \Lambda_{n,i} w_{n,i}}{\mu_n(l)} < \infty . \tag{3.2.6}$$

The ergodicity condition has an interesting interpretation. If at all resources $n = 1,...,N$

$$\liminf_{k \to \infty} \mu_n(k) > \sum_{i \in O} \Lambda_{n,i} w_{n,i} , \tag{3.2.7}$$

the ergodicity condition is satisfied. The condition states that the total amount of work offered to a resource n must not exceed the capacity of the resource.
Note that this condition is very similar to the ergodicity condition of a single Markovian queue with a Poisson arrival process and queue length dependent service rates.

The separability condition.

The limiting distribution has a product form if for all states $s \in S(\boldsymbol{K})$ and at all resources $n = 1,...,N$

$$\sum_{k=1}^{k_n} \frac{\phi_n(k,k_n) - \gamma_n(k,k_n)}{w_{n,r_n(k)}} = 0 . \tag{3.2.8}$$

The consequences of this condition are discussed later on.

We now may formulate the main theorem of this section.

Theorem 3.1

If the Markov process $\{X(t), t \geqslant 0\}$ satisfies the ergodicity and separability conditions, the limiting and stationary probabilities $p(s,\boldsymbol{K})$ are for $s \in S(\boldsymbol{K})$ given by

$$p(s,\boldsymbol{K}) = \frac{\pi(s)}{G(\boldsymbol{K})} , \tag{3.2.9}$$

where

$$\pi(s) = \prod_{n=1}^{N} F_n(s_n) \tag{3.2.10}$$

and

$$G(\boldsymbol{K}) = \sum_{s \in S(\boldsymbol{K})} \pi(s) ; \tag{3.2.11}$$

the quantity $F_n(s_n)$ equals

$$F_n(s_n) = \prod_{k=1}^{k_n} \frac{x_{n,r_n(k)}}{\mu_n(k)} , \tag{3.2.12}$$

where for $r \in C$ equals

$$x_{n,r} = f_{n,r} w_{n,r} \tag{3.2.13}$$

and for $i \in O$

$$x_{n,i} = \Lambda_{n,i} w_{n,i} . \tag{3.2.14}$$

∎

This theorem has far reaching consequences as it allows for the construction of relatively efficient procedures for the computation of steady-state system characteristics.

Let us discuss some restrictions and possible extensions of the separability condition.

The following two lemmas give straightforward conditions for the separability condition to hold.

Lemma 3.1
The separability condition is satisfied at resource n if there is a value w_n such that for all $r \in C \cup O$

$$w_n = w_{n,r} \ . \tag{3.2.15}$$

Proof: Verification. ∎

Lemma 3.2
The separability condition is satisfied at resource n if the dedication and admittance function are such that for all $K = 1,2,...$ and $k = 1,...,K$

$$\phi_n(k,K) = \gamma_n(k,K) \tag{3.2.16}$$

Proof: Verification. ∎

The service disciplines which satisfy (3.2.16) are called symmetric disciplines and play an important role in the analysis of separable queueing network models, cf. Kelly [1976] and [1979] and Hordijk and Van Dijk [1981]. The processor sharing and infinite server disciplines are such symmetric disciplines.

The admittance and dedication function of the first-come first-served discipline does not satisfy this condition. So, it must be assumed that at such resources (3.2.15) is satisfied, i.e. at each first-come first-served resource the customers must have exponentially distributed service demands with a common mean.

It should be observed that a rather simple routing behaviour and exponentially distributed service demands have been introduced. It is well known that these restrictions may be relaxed somewhat, cf. Baskett, Chandy, Muntz and Palacios [1975], Kelly [1976] and [1979] and Hordijk and Van Dijk [1981].
This may be done by the introduction of class indices allowing for the introduction of a more detailed description of the routing behaviour. Apart from a chain number each customer attains a class index that, in contrast with the chain number, may change during its sojourn in the system.
Class indices can be used to represent (a part of the) history of the route that a specific customer has followed. As a consequence routing behaviour and service demands can be made history dependent. This makes it possible to use class indices as a tool to simulate non-exponential service demand distributions at resources with a symmetric service discipline.
In fact, it can be shown that the limiting probabilities attain exactly the same product

form as sketched in Theorem 3.1, if the service demand of a customer of chain r at a resource n with a symmetric service discipline has a distribution function with a rational Laplace-Stieltjes transform instead of an exponential distribution.

3.3. The MVA approach

3.3.1. Introduction

In the preceding section a class of mixed multichain queueing network models has been introduced and the product form solution has been formulated for the limiting distribution of a relatively detailed state description under certain ergodicity and separability conditions.
A normalization constant appears in the denominator of the expression for the limiting probabilities. This constant may, in principle, be evaluated by a summation of relatively complicated expressions over the complete state space. Obviously, such a technique has to be dissuaded as the state space of a queueing network model can be enormous.
Moreover, the interest is not so much in these limiting probabilities for a detailed description of the buffer occupation, as well as in more global steady state characteristics. In the literature two elegant and relatively efficient approaches for the construction of computational algorithms have been introduced: the convolution method and the mean value analysis. In this monograph we concentrate at the second approach.

The MVA approach is based on a set of recursive relations between certain steady state characteristics. In Reiser [1979:1] the approach has been presented for a simple queueing network with first-come first-served resources with a single service unit and one closed customer chain. It was observed that the relations might be interpreted in terms of an arrival theorem and Little's formula.
Afterwards, the approach has been extended to closed multichain queueing network models in Reiser and Lavenberg [1980], and mixed open and closed multichain queueing network models in for instance Zahorjan and Wong [1981], Krzesinski, Teunissen and Kritzinger [1982] and Bruell, Balbo and Afshari [1984].
Though in all papers the interpretation in terms of an arrival theorem and Little's formula is referred to, this interpretation is not given. In this section we, therefore, shall sketch this interpretation for the mixed open and closed queueing network model satisfying the ergodicity and separability conditions.
In the next section we discuss the implementation of the relations in the MVA algorithm.

In Subsection 3.3.2 we introduce important steady state system characteristics as expectations of the corresponding limiting processes. The relation with the stationary versions of these processes is indicated. The subsection is concluded with a theorem summarizing the mean value relations between the steady state characteristics.
In Subsection 3.3.3 Little's formula and an arrival theorem are discussed and some first applications are presented.

In Subsection 3.3.4 the MVA approach is presented for queueing network models with queue length dependent service rate functions. We start with the discussion of an open and a closed single chain queueing network model. The reasoning is then extended to the general case.
In Subsection 3.3.5 the MVA approach is presented for queueing network models with fixed service rate functions. It is shown that for such networks one may concentrate at the evaluation of closed multichain queueing network models with an adjusted set of service demand parameters.

3.3.2. Steady state system characteristics and mean value relations

To describe the steady state system characteristics, for all $r \in C \cup O$, $n \in Q$, $k = 1,2,...$, $t \geqslant 0$, $\nu = 1,2,...$ and $h > 0$ the following random variables are introduced

$S_{n,r}(\nu,K)$	the residence time at resource n of the νth arriving customer of chain r,
$A_{n,r}(t,t+h,K)$	the number of customers of chain r that arrive at resource n in the interval $[t,t+h)$,
$D_{n,r}(t,t+h,K)$	the number of customers of chain r that depart from resource n in the interval $[t,t+h)$,
$L_{n,r}(t,K)$	the number of customers of chain r at resource n at time t, and
$I_n(t,k,K)$	the indicator function that at time t attains the value 1 if there are k customers at resource n and the value 0 otherwise.

The argument K emphasizes the dependence on the populations of the closed customer chains.

If the queueing network model satisfies the ergodicity conditions, the limiting distributions of these random variables exist and we may introduce $S_{n,r}(K)$, $\Lambda_{n,r}(K)$, $L_{n,r}(K)$ and $p_n(k,K)$ as

$$S_{n,r}(K) = \lim_{\nu \to \infty} E\{S_{n,r}(\nu,K)\}, \tag{3.3.1}$$

$$\Lambda_{n,r}(K) = \lim_{h \downarrow 0} \lim_{t \to \infty} \frac{E\{A_{n,r}(t,t+h,K)\}}{h} = \lim_{h \downarrow 0} \lim_{t \to \infty} \frac{E\{D_{n,r}(t,t+h,K)\}}{h} \tag{3.3.2}$$

$$L_{n,r}(K) = \lim_{t \to \infty} E\{L_{n,r}(t,K)\}, \text{ and} \tag{3.3.3}$$

$$p_n(k,K) = \lim_{t \to \infty} E\{I_n(t,k,K)\}. \tag{3.3.4}$$

We can go one step further by appealing to the fact that the process $\{X(t), t \geqslant 0\}$ is a regenerative process. Note that every initial state may be taken as a reference point. If the ergodicity conditions are satisfied, the expected length of a regeneration cycle will be

finite.
As a consequence, cf. Heyman and Sobel [1982:pp 362-380], we can study the steady state behaviour by an analysis of the stationary version of the stochastic process. This greatly facilitates the evaluation of steady state characteristics.

Assuming $\{X(t), t \geqslant 0\}$ to be stationary we may write

$$S_{n,r}(K) = E\{S_{n,r}(1,K)\},$$

$$\Lambda_{n,r}(K) = \lim_{h \downarrow 0} \frac{E\{A_{n,r}(0,h,K)\}}{h} = \lim_{h \downarrow 0} \frac{E\{D_{n,r}(0,h,K)\}}{h},$$

$$L_{n,r}(K) = E\{L_{n,r}(0,K)\},$$

$$p_n(k,K) = E\{I_n(0,k,K)\}.$$

We use the following formulations for the steady state or stationary behavioural characteristics

$S_{n,r}(K)$	expected residence time (including waiting and service) of a customer of chain r at resource n,
$\Lambda_{n,r}(K)$	throughput of customers of chain r at resource n,
$L_{n,r}(K)$	expected number of customers of chain r at resource n and
$p_n(k,K)$	probability that k customers are at resource n.

In the literature a great number of relations between these steady state characteristics have been formulated and proved. Before the MVA approach is presented, we summarize the relations that are relevant for our purposes in the form of a theorem. As we have noted these relations can be proved by algebraic manipulation of the detailed limiting probabilities. The proofs fall outside the scope of this monograph. For similar results and details on the method of proof we refer to Reiser and Lavenberg [1980], Krzesinski, Teunissen and Kritzinger [1982] and Bruell, Balbo and Afshari [1984]. The last paper contains an overview of known relations and references are provided for a more detailed study.

Theorem 3.2: The Mean Value Relations

Let $M_n(K)$, $n \in Q$, denote the maximum number of customers that may be present at resource n, if the population vector is K, and let the buffer occupation process $\{X(t), t \geqslant 0\}$ satisfy the ergodicity and separability conditions.

Then, for a closed chain $r \in C$ the characteristics at resource $n \in Q$ are related through

$$S_{n,r}(K) = \sum_{k=1}^{M_n(K)} \frac{k w_{n,r}}{\mu_n(k)} p_n(k-1, K-e_r), \tag{3.3.5}$$

$$\Lambda_{n,r}(K) = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(K)}, \tag{3.3.6}$$

$$L_{n,r}(K) = \Lambda_{n,r}(K) S_{n,r}(K) , \tag{3.3.7}$$

and for an open chain $i \in O$ through

$$S_{n,i}(K) = \sum_{k=1}^{M_n(K)} \frac{k w_{n,i}}{\mu_n(k)} p_n(k-1,K) , \tag{3.3.8}$$

$$\Lambda_{n,i}(K) = \Lambda_{n,i} , \tag{3.3.9}$$

$$L_{n,i}(K) = \Lambda_{n,i} S_{n,i}(K) . \tag{3.3.10}$$

The probabilities $p_n(k,K)$ are, for $k = 1,...,M_n(K)$, related through the difference equations

$$p_n(k,K) = \sum_{i \in O} \frac{w_{n,i} \Lambda_{n,i} p_n(k-1,K)}{\mu_n(k)} + \sum_{r \in C} \frac{w_{n,r} \Lambda_{n,r}(K) p_n(k-1,K-e_r)}{\mu_n(k)} \tag{3.3.11}$$

and a boundary condition

$$\sum_{k=0}^{M_n(K)} p_n(k,K) = 1 . \tag{3.3.12}$$

■

The mean value relations form a recursive scheme for the computation of the introduced performance characteristics. In Section 3.4 this scheme is studied in more detail. The rest of this section is devoted to an interpretation of the relations.

3.3.3. Little's formula and an arrival theorem

In this subsection we discuss two important tools in the analysis of separable queueing network models: Little's formula and an arrival theorem.

Little's formula, cf. Little [1961], Stidham [1974], Brumelle [1971] and Heyman and Sobel [1982], relates for a stationary queueing system the expected number of customers, the rate of the arrival process and the expected residence time of a customer as

$$L = \lambda S . \tag{3.3.13}$$

The virtue of Little's formula is its wide applicability which is mainly due to the flexibility one has in choosing "the queueing system" to which $L = \lambda S$ is applied.
To illustrate its use in the analysis of queueing network systems we shall discuss a derivation of the Relations (3.3.7) and (3.3.10) and (3.3.6). It is assumed that the buffer occupation process of Subsection 3.2.3 satisfies the ergodicity conditions. So, we can use the stationary version of the buffer occupation process. Note that we do not demand that the separability conditions are satisfied.

The first application is rather straightforward. Concentrating on the analysis of the characteristics of customers of a certain closed chain $r \in C$ at a resource $n \in Q$, we observe that the stochastic processes $\{L_{n,r}(t,K), t \geqslant 0\}$ and $\{S_{n,r}(v,K), v=1,2,...\}$ are stationary and that the arrival rate is $\Lambda_{n,r}(K)$. A direct application of Little's formula yields (3.3.7).
An analogous reasoning may be used to ascertain (3.3.10).

A more sophisticated application yields (3.3.6). We concentrate on the analysis of the system characteristics of a certain closed chain $r \in C$ at a particular resource n. Relation (3.3.6) can be rewritten as

$$K_r = \Lambda_{n,r}(K) \sum_{m=1}^{N} \frac{f_{m,r}}{f_{n,r}} S_{m,r}(K) .$$

It will be made plausible that this relation is an application of Litle's formula with the complete queueing network system as "the queueing system".
Assume that the customers of chain r arriving at resource n are observed. First of all, the arrival rate equals $\Lambda_{n,r}(K)$. If we can distinguish between the customers of chain r, the time between two arrivals at resource n can be measured for each individual customer. With $f_{m,r}/f_{n,r}$ being the expected number of visits to resource m between two visits to resource n the expected time of such a round trip through the system is

$$\sum_{m=1}^{N} \frac{f_{m,r}}{f_{n,r}} S_{m,r}(K) .$$

The expected number of customers of chain r in the system equals K_r. Applying Little's formula we find the desired result (3.3.13).

Note that these two illustrations of the use of Little's formula have a much wider applicability than just the separable mixed open and closed queueing network system introduced in Section 3.2.

The second important property of queueing network systems applies typically to separable queueing networks. For mixed multichain queueing network models that satisfy the ergodicity and separability conditions, one may formulate a relation between the limiting distribution of the queueing processes and the limiting distribution of the queueing processes at arrival, jump and departure moments.
The main result is a so-called arrival theorem which, roughly spoken, states that an arbitrary customer observes at an arrival, jump or departure moment the queueing network system as if in equilibrium with itself removed. For a customer of an open chain this implies that it observes the system in equilibrium and for a customer of a closed chain that it observes equilibrium as if one customer of its own chain has been removed from the system.
For closed multichain queueing networks an arrival theorem has been stated and proved in Lavenberg and Reiser [1982], cf. also Kelly [1979:Theorem 3.12]. For open multichain queueing networks some results have been formulated in Kelly [1979:Theorem 3.7].

For an open chain i, $i \in O$ the following stationary probabilities for the queueing network model with population vector $\boldsymbol{K}$ are introduced for all $s \in S(\boldsymbol{K})$ and $n, m \in Q$

$a_{i,n}(s, \boldsymbol{K})$ stationary probability that a customer of an open chain i arriving at resource n, sees the system in state s,

$q_{i,n,m}(s, \boldsymbol{K})$ statinary probability that a customer of an open chain i jumping from resource n to resource m, sees the system in state s,

$d_{i,n}(s, \boldsymbol{K})$ stationary probability that a customer of an open chain i departing from the system after its service has been completed at resource n, sees the system in state s,

For a closed chain $r \in C$ the following stationary probabilities are introduced for all $s \in S(\boldsymbol{K}-e_r)$ and $n, m \in C$

$q_{r,n,m}(s, \boldsymbol{K})$ stationary probability that a customer of chain r jumping from resource n to resource m, sees the system in state s.

We now may formulate the following theorem.

Theorem 3.3 : The Arrival Theorem

Consider the stationary version of the buffer occupation process $\{X(t), t \geqslant 0\}$ of Section 3.2.3 satisfying the ergodicity and separability conditions.
Then for all $i \in O$, $n, m = 1,\ldots,N$, and $s \in S(\boldsymbol{K})$

$$a_{i,n}(s, \boldsymbol{K}) = q_{i,n,m}(s, \boldsymbol{K}) = d_{i,n}(s, \boldsymbol{K}) = \frac{\pi(s)}{G(\boldsymbol{K})}, \tag{3.3.14}$$

and for all $r \in C$, $n, m = 1,\ldots,N$, and $s \in S(\boldsymbol{K}-e_r)$

$$q_{r,n,m}(s, \boldsymbol{K}) = \frac{\pi(s)}{G(\boldsymbol{K}-e_r)} \tag{3.3.15}$$

Proof :

We proof Relation (3.3.15). The other relations may be proved in a similar way.

We introduce $A_{r,n,m}(s)$ as the rate of the stationary process describing the number of jumps of customers from a closed chain r from resource n to resource m observing the system at the jump moment in a state s, where $r \in C$, $n, m = 1,\ldots,N$ and $s \in S(\boldsymbol{K}-e_r)$.

To describe a change in a state vector s a shift operator $T_{n,k,r}(s)$ is introduced, where $n = 1,\ldots,N$, $k = 1,2,\ldots$ and $r = 1,\ldots,L$. The shift operator affects the buffer description of resource n by the insertion of a customer of chain r at buffer-place k. So, if the nth component of s equals

$$s_n = (k_n, r_n(1), \ldots, r_n(k_n)),$$

the nth component of $T_{n,k,r}(s)$ equals

$$\left[T_{n,k,r}(s)\right]_n = (k_n+1, r_n(1),..,r_n(k-1), r, r_n(k),..,r_n(k_n))$$

Applying the results of Theorem 3.1 we obtain for those $s \in S(K-e_r)$ for which $T_{n,k,r}(s) \in S(K)$

$$\begin{aligned} A_{r,n,m}(s) &= \sum_{k=1}^{k_{n+1}} p(T_{n,k,r}(s),K)\ \frac{\mu_n(k_n+1)\phi_n(k,k_n+1)P_{r,n,m}}{w_{n,r}} \\ &= \sum_{k=1}^{k_n+1} \frac{\pi(s)}{G(K)}\ \frac{f_{n,r}w_{n,r}}{\mu_n(k_n+1)}\ \frac{\mu_n(k_n+1)\phi_n(k,k_n+1)P_{r,n,m}}{w_{n,r}} \\ &= \frac{f_{n,r}P_{r,n,m}}{G(K)}\ \pi(s) \end{aligned}$$

and consequently, referring to Relation (3.2.9),

$$\begin{aligned} q_{r,n,m}(s) &= \frac{A_{r,n,m}(s)}{\sum_{u \in S(K-e_r)} A_{r,n,m}(u)} \\ &= \frac{f_{n,r}P_{r,n,m}\ \pi(s)}{f_{n,r}P_{n,r,m} \sum_{u \in S(K-e_r)} \pi(u)} \\ &= p(s, K-e_r) \end{aligned}$$

∎

In the next subsections we shall see some applications of the arrival theorem.

3.3.4. The MVA approach: queue length dependent service rates

3.3.4.1. Introduction

In this subsection we sketch the MVA approach for the queueing network model satisfying the conditions for ergodicity and separability. The approach provides an intuitive derivation for the mean value relations stated in Theorem 3.2.
To illustrate the line of reasoning we start with two special cases. In Subsections 3.3.4.2 and 3.3.4.3 the approach is highlighted for an open and a closed queueing network with one customer chain, respectively. Here, the reasonings can be seen as informal proofs.
In Subsection 3.3.4.3 we discuss the general case with multiple open and closed customer chains. Here, the reasoning provides an intuitive understanding only.

3.3.4.2. An open single chain queueing network model

Consider a queueing network system with one open customer chain. There are no closed customer chains. So both the chain number and the argument K may be dropped.

The model is described by the following parameters

N	number of resources,
$\mu_n(k)$	service rate of resource n if k customers are present, and
w_n	mean service demand at resource n.

The steady state system characteristics are

S_n	expected residence time at resource n,
Λ_n	throughput of customers at resource n,
L_n	expected number of customers at resource n, and
$p_n(k)$	probability that k customers are at resource n.

The MVA approach starts with an intuitive derivation of relations for the probabilities $p_n(k)$.
It is common practice to the derive the limiting distribution of a one-dimensional Markovian queueing system or random walk by matching the flows between two adjacent states in the one-dimensional state space.
This reasoning shall be applied to a particular resource n, $n = 1,\ldots,N$. Due to the arrival theorem an arriving customer sees $k-1$ customers present with probability $p_n(k-1)$, where $k = 1,2,\ldots$. The throughput at resource n equals Λ_n. So, the average number of transitions from state $k-1$ to state k per unit time equals $\Lambda_n p_n(k-1)$.
On the other hand the resource operates at a rate $\mu_n(k)$ if k customers are present and so, the average number of transitions from state k to state $k-1$ per unit time equals $w_n^{-1}\mu_n(k)p_n(k)$. Matching these two flows yields, for $k = 1,2,\ldots$,

$$\frac{\mu_n(k)}{w_n} p_n(k) = \Lambda_n p_n(k-1) . \tag{3.3.16}$$

A boundary condition follows from the fact that the marginal queue length distribution must be proper, i.e.

$$\sum_{k=0}^{\infty} p_n(k) = 1 . \tag{3.3.17}$$

The Relations (3.3.16) and (3.3.17) are special cases of (3.3.11) and (3.3.12) respectively. The solution is, for $k = 0,1,\ldots$, given by

$$p_n(k) = \prod_{l=1}^{k} \frac{w_n \Lambda_n}{\mu_n(l)} \left[1 + \sum_{j=1}^{\infty} \prod_{l=1}^{j} \frac{w_n \Lambda_n}{\mu_n(l)} \right]^{-1} , \tag{3.3.18}$$

where the empty product is defined to be one.

Note that this marginal queue length distribution forms a proper distribution function if

$$1 + \sum_{k=1}^{\infty} \prod_{l=1}^{k} \frac{w_n \Lambda_n}{\mu_n(l)} < \infty . \tag{3.3.19}$$

This condition agrees with the ergodicity condition (3.2.6). Note that the marginal queue length distribution equals the stationary distribution of a one-dimensional Markovian queue with a Poisson arrival process with rate Λ_n, queue length dependent service rates $\mu_n(k)$ and exponential service demands with mean w_n.

Rewriting (3.3.16) as

$$\mu_n(k)p_n(k) = w_n \Lambda_n p_n(k-1) , \tag{3.3.20}$$

we may give another interpretation that shall be of importance in the analysis of the general queueing network model. The left hand side of this relation denotes the rate at which work is done with k customers being present, whereas the right hand side denotes the rate at which work is arriving with $k-1$ customers being present. Apparently, these rates are equal for this separable queueing network model. We shall refer to these rates as "work flow rates".

The next step in the MVA approach is the derivation of a relation for the expected residence time at resource n. Multiplying both sides of (3.3.16) by k and summing over $k = 1,2,...$ yields

$$\sum_{k=1}^{\infty} kp_n(k) = \Lambda_n \sum_{k=1}^{\infty} \frac{kw_n}{\mu_n(k)} p_n(k-1) . \tag{3.3.21}$$

Observe that the left hand side of (3.3.21) equals the expected number of customers at resource n. Application of Little's formula yields

$$S_n = \sum_{k=1}^{\infty} \frac{kw_n}{\mu_n(k)} p_n(k-1) . \tag{3.3.22}$$

This relation corresponds with (3.3.8).

For the expected number of customers at resource n we have by virtue of Little's formula

$$L_n = \Lambda_n S_n . \tag{3.3.23}$$

This completes the MVA approach. A computational algorithm for the evaluation of the characteristics is relatively simple to design as the resources of the queueing network model may be analysed in isolation. The only coupling is by equation (3.2.3) which relates the throughputs at the different resources. If it is possible to simplify the evaluation of (3.3.18), the amount of work to be done will be very small. In other cases the algorithm involves a summation over an infinite number of terms.

3.3.4.3. A closed single chain queueing network

In this subsection we present the MVA approach for a closed single chain queueing network model with queue length dependent service rate functions. So, we may omit the chain number in the formulations.
The system is described by the following parameters

N	number of resources,
$\mu_n(k)$	service rate of resource n if k customers are present,
K	number of customers in the system,
w_n	mean service demand of a customer at resource n, and
f_n	visiting ratio at resource n.

The steady state characteristics are

$S_n(K)$	expected residence time at resource n,
$\Lambda_n(K)$	throughput at resource n,
$L_n(K)$	expected number of customers at resource n, and
$p_n(k,K)$	marginal probability that k customers are at resource n.

The MVA approach starts with a derivation of the marginal queue length distributions at a particular resource n, $n=1,\ldots,N$.
Due to the arrival theorem an arriving customer will see $k-1$ customers present with probability $p_n(k-1,K-1)$. The throughput is $\Lambda_n(K)$. So, the transitions rate from state $k-1$ to state k equals $\Lambda_n(K)p_n(k-1,K-1)$. On the other hand, the resource operates at a rate $\mu_n(k)$ with k customers being present and so, the transition rate from state k to state $k-1$ equals $w_n^{-1}\mu_n(k)p_n(k,K)$. Matching these two transition rates we obtain for $k=1,\ldots,K$

$$\frac{\mu_n(k)}{w_n}\,p_n(k,K) = \Lambda_n(K)p_n(k-1,K-1)\,. \tag{3.3.24}$$

The boundary condition is

$$\sum_{k=0}^{K} p_n(k,K) = 1\,. \tag{3.3.25}$$

These relations imply a simple recursion, once the throughput $\Lambda_n(K)$ is known. Note that here an interpretation in terms of work flow rates can be given as well, cf. Relation (3.3.20).

The next step in the MVA approach is the derivation of a relation for the expected residence times at resource n. Multiplying both sides of (3.3.24) by k and summing over $k=1,\ldots,K$ yields

$$\sum_{k=1}^{K} kp_n(k,K) = \Lambda_n(K)\sum_{k=1}^{K}\frac{kw_n}{\mu_n(k)}\,p_n(k-1,K-1)\,. \tag{3.3.26}$$

We recognise the left hand side as the expected number of customers at resource n and find by application of Little's formula

$$S_n(K) = \sum_{k=1}^{K} \frac{k w_n}{\mu_n(k)} p_n(k-1, K-1) . \tag{3.3.27}$$

For a derivation of a throughput relation at resource n we refer to the discussion of Little's formula in 3.3.2, i.e.

$$\Lambda_n(K) = \frac{f_n K}{\sum_{m=1}^{N} f_m S_m(K)} . \tag{3.3.28}$$

The last step in the MVA approach is another application of Little's formula. At resource n we obtain for the expected number of customers

$$L_n(K) = \Lambda_n(K) S_n(K) . \tag{3.3.29}$$

Starting with $L_n(0)=0$ at all resources $n=1,...,N$ the Relations (3.3.27), (3.3.28), (3.3.24), (3.3.25) and (3.3.29) form a finite and recursive scheme for the evaluation of the steady state characteristics of the queueing network model.

3.3.4.4. Mixed multichain queueing network models

In this subsection the mixed open and closed multichain queueing network model that satisfies the ergodicity and separability conditions is considered. The ideas sketched in the preceding subsections are used to give an intuitive derivation of the mean value relations. We first describe the parameters of the system and the characteristics that are to be evaluated. Then the MVA approach is sketched.

The model is characterised by the following parameters

N	number of resources,
$\mu_n(k)$	service rate at resource n if k customers are present,
O	set of open chains,
C	set of closed chains,
K	population vector of closed customer chains,
$f_{n,r}$	visiting ratio of closed chain r customers at resource n,
$M_n(K)$	maximum number of customers that may be present at resource n.

The following steady state characteristics play a role in the analysis

$S_{n,r}(K)$	expected residence time of chain r customers at resource n,
$\Lambda_{n,i}$	throughput of an open chain i at resource n,

$\Lambda_{n,r}(K)$ throughput of chain r customer at resource n,

$L_{n,r}(K)$ expected number of chain r customers at resource n,

$p_n(k,K)$ probability that k customers are at resource n.

The first step in the MVA approach concerns the derivation of the Relations (3.3.11) and (3.3.12) for the marginal queue length distribution at a particular resource n.
We cannot use the technique of matching the transition rates between two states $k-1$ and k directly. But applying the interpretation in terms of a "work flow rate" between these two states, we can give an intuition for Relation (3.3.11).
The work flow rate from state k to state $k-1$ equals for all $k=1,...,M_n(K)$

$$\mu_n(k)p_n(k,K).$$

For the work flow rate from state $k-1$ to state k we must distinguish between the open and closed customer chains. The customers of an open chain $i \in O$ arrive at a rate $\Lambda_{n,i}$ and have a mean service demand of $w_{n,i}$. Due to the arrival theorem an arriving customer sees $k-1$ customers present with probability $p_n(k-1,K)$. So, the work flow rate from state $k-1$ to k due to work from chain i equals

$$w_{n,i}\Lambda_{n,i}p_n(k-1,K).$$

For the work flow rate due to work from a closed chain $r \in C$ we thus find with the same reasoning

$$w_{n,r}\Lambda_{n,r}(K)p_n(k-1,K-e_r).$$

Combining these results we find for all $k=1,....,M_n(K)$

$$\mu_n(k)p_n(k,K) = \sum_{i \in O} w_{n,i}\Lambda_{n,i}p_n(k-1,K) + \sum_{r \in C} w_{n,r}\Lambda_{n,r}(K)p_n(k-1,K-e_r) \tag{3.3.30}$$

with as boundary condition

$$\sum_{k=0}^{M_n(K)} p_n(k,K) = 1 \tag{3.3.31}$$

Relations (3.3.30) and (3.3.31) correspond with (3.3.11) and (3.3.12) respectively.

The second step in the MVA approach is the derivation of a relation for the expected residence time of a customer of a particular chain at a specific resource.
Multiplying both sides of (3.3.30) by k and summing over $k=1,...,M_n(K)$ yields

$$\sum_{k=1}^{M_n(K)} kp_n(k,K) = \sum_{i \in O} \Lambda_{n,i} \sum_{k=1}^{M_n(K)} \frac{kw_{n,i}}{\mu_n(k)} p_n(k-1,K) + \sum_{r \in C} \Lambda_{n,r}(K) \sum_{k=1}^{M_n(K)} \frac{kw_{n,r}}{\mu_n(k)} p_n(k-1,K-e_r). \tag{3.3.32}$$

The left hand side of (3.3.32) is the expected numbers of customers at resource n, so

$$\sum_{k=1}^{M_n(K)} k p_n(k,K) = \sum_{i \in O} L_{n,i}(K) + \sum_{r \in C} L_{n,r}(K). \tag{3.3.33}$$

Observe that by virue of Little's formula for all $i \in O$

$$L_{n,i}(K) = \Lambda_{n,i} S_{n,i}(K), \tag{3.3.34}$$

and for all $r \in C$

$$L_{n,r}(K) = \Lambda_{n,r}(K) S_{n,r}(K). \tag{3.3.35}$$

Combining the Relations (3.3.32), (3.3.34) and (3.3.35), it seems a reasonable guess that for all $r \in C$

$$S_{n,r}(K) = \sum_{k=1}^{M_n(k)} \frac{k w_{n,r}}{\mu_n(k)} p_n(k-1, K-e_r) \tag{3.3.36}$$

and for all $i \in O$

$$S_{n,i}(K) = \sum_{k=1}^{M_n(k)} \frac{k w_{n,i}}{\mu_n(k)} p_n(k-1, K). \tag{3.3.37}$$

Relations (3.3.36) and (3.3.37) agree with (3.3.5) and (3.3.8) respectively. Note that this result suggests that there is a work flow balance per customer chain.

The third step in the MVA approach is the derivation of the throughputs of a closed chain $r \in C$ at a particular resource n. Referring to the discussion in Subsection 3.3.2 we find

$$\Lambda_{n,r}(K) = \frac{f_{n,r} K_r}{\sum_{m=1}^{M} f_{m,r} S_{m,r}(K)}. \tag{3.3.38}$$

Note that (3.3.38) agrees with (3.3.6).

Thus we have given an intuitive derivation of the relations between the system characteristics as formulated in in Theorem 3.3. In the next section we show that these relations may be used to design a recursive algorithm for the evaluation of the characteristics at arbitrary population vectors.

3.3.5. The MVA approach: fixed service rate functions

In this subsection the MVA approach is discussed for a class of mixed multichain queueing network models with resources falling in of the following three categories:

FCFS — A resource with a first-come first served service discipline and a single service unit operating atconstant service rate. The service demands of the distinct customer chains are independent and exponentially distributed with a common mean.

PS A resource with a processor sharing service discipline and a single service unit operating a constant service rate. The service demands of the different customer chains are independent and exponentially distributed.

IS A resource with an infinite server service discipline. The service rate of the service units is constant. The service demands of the different customer chains are independent and exponentially distributed.

For this class of queueing network models the mean value relations attain an attractive form allowing for the efficient evaluation of the main system characteristics. It appears that expected residence times, throughputs and expected numbers of customers can be evaluated from a relatively simple recursive scheme. The evaluation of the marginal queue length distributions is not necessary in this case.

The model is characterized by the following parameters

N number of resources,

μ_n service rate of a single service unit at resource n,

O set of open customer chains,

C set of closed customer chains,

K population vector of closed customer chains,

$f_{n,r}$ visiting ratio of customers of a closed chain r at resource n,

w_n mean service demand at a FCFS resource n,

$w_{n,r}$ mean service demand of a customer of chain r at PS or IS resource n.

The relevant steady state characteristics are

$S_{n,r}(K)$ expected residence time of a customer of chain r at resource n,

$\Lambda_{n,i}$ throughput of customers of an open chain i at resource n

$\Lambda_{n,r}(K)$ throughput of customers of a closed chain r at resource n,

$L_{n,r}(K)$ expected number of customers of chain r at resource n.

The reasoning for the derivation of a relation for the throughputs and expected numbers of customers by application of Little's formula may be repeated for this case. We thus immediately obtain for all chains $r \in C$ at the resources $n = 1,...,N$

$$\Lambda_{n,r}(K) = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(K)} \tag{3.3.40}$$

and

$$L_{n,r}(K) = \Lambda_{n,r}(K) S_{n,r}(K) , \tag{3.3.41}$$

and for all chains $i \in O$ at the resources $n = 1,...,N$

$$\Lambda_{n,i}(K) = \Lambda_{n,i} \tag{3.3.42}$$

and

$$L_{n,i}(K) = \Lambda_{n,i} S_{n,i}(K) . \tag{3.3.43}$$

It will be shown that the MVA approach leads to an appealing interpretation of relatively simple expressions for the expected residence times. The reasoning is different for the different types of service disciplines. So, we shall consider resources with FCFS, PS and IS service disciplines respectively.

Let n be a resource with a FCFS service discipline.

As a consequence of the arrival theorem, a customer of an open chain i, $i \in O$, sees on the average $L_{n,l}(K)$ customers of chain l, $l \in O \cup C$, in front of it. The expected remaining service demand equals for all these customers w_n. So, the expected residence time of a customer of an open chain i at the FCFS resource n is

$$S_{n,i}(K) = (\sum_{l \in O \cup C} L_{n,l}(K) + 1) \frac{w_n}{\mu_n} . \tag{3.3.44}$$

As a consequence of the arrival theorem, a customer of a closed chain r, $r \in C$, sees on the average $L_{n,l}(K - e_r)$ customers of chain l, $l \in O \cup C$, in front of it. The expected residence time of the customer of a closed chain r thus is

$$S_{n,r}(K) = (\sum_{l \in O \cup C} L_{n,l}(K - e_r) + 1) \frac{w_n}{\mu_n} . \tag{3.3.45}$$

These relations follow also from (3.3.8) and (3.3.5) by inserting $\mu_n(k) = \mu_n$, $k = 1,2,...$ and $w_{n,l} = w_n$, for all $l \in O \cup C$.

Let n be a resource with a PS service discipline.

As a consequence of (3.3.8) and (3.3.5) the expected residence time of a customer of an open chain i, $i \in O$, at resource n equals

$$S_{n,i}(K) = (\sum_{l \in O \cup C} L_{n,l}(K) + 1) \frac{w_{n,i}}{\mu_n} \tag{3.3.46}$$

and of a customer of a closed chain r, $r \in C$,

$$S_{n,r}(K) = (\sum_{l \in O \cup C} L_{n,l}(K - e_r) + 1) \frac{w_{n,r}}{\mu_n} . \tag{3.3.47}$$

These relations have an interesting intuitive explanation. The arrival theorem states that a customer of an open chain observes the system upon an arrival, jump or departure moment as if in equilibrium. If the system is in equilibrium during the residence of this

particular customer, the expected residence time equals the product of the mean rate at which the service unit works for this customer and the mean service demand, i.e.

$$S_{n,i}(K) = \left[\frac{1}{\sum_{l \in O \cup C} L_{n,l}(K) + 1} \right]^{-1} \frac{w_{n,i}}{\mu_n} . \tag{3.3.48}$$

For a customer of a closed chain r this argument yields

$$S_{n,r}(K) = \left[\frac{1}{\sum_{l \in O \cup C} L_{n,l}(K - e_r) + 1} \right]^{-1} \frac{w_{n,r}}{\mu_n} . \tag{3.3.49}$$

Let n be a resource with a IS service discipline.

Each arriving customer is served immediately and consequently the expected residence time equals for all customers of an open or closed chain l, $l \in O \cup C$

$$S_{n,l}(K) = \frac{w_{n,l}}{\mu_n} . \tag{3.3.50}$$

In the next Section we will see that the above introduced relations form the basis of an efficient algorithm for evaluating the main system characteristics.

3.4. The MVA algorithm

3.4.1. Introduction

Relations (3.3.5) through (3.3.12) form the basis of a recursive algorithm for the computation of the introduced system characteristics. This algorithm is called the Mean Value Analysis (MVA) algorithm. This Section is concerned with the formulation and implementation of the algorithm and is organized as follows.

In Subsection 3.4.2 the MVA algorithm is discussed for general mixed multichain queueing network models. It appears that complicated expressions have to be evaluated and that extra conditions have to be imposed to allow for an efficient computational procedure.
In Subsection 3.4.3 we discuss the MVA algorithm for the network model treated in Subsection 3.3.5. For this special case the MVA algorithm takes a simple.
In Subsection 3.4.4 we discuss the implementation of the recursion and make some remarks with respect to the computational complexity of the resulting algorithm.

3.4.2. The MVA algorithm: queue length dependent service rates

The mean value relations form the basis of a recursive algorithm for the evaluation of system characteristics in mixed multichain queueing network models. The Relations (3.3.5) and (3.3.11) indicate that the algorithm is recursive in the population vector. To

compute the characteristics at population vector K we need the characteristics at the population vectors $K-e_r$ for all $r \in C$.
So, the recursion runs through all vectors in the range of (0,...,0) through $(K_1, \ldots, K_R)$ in order to compute the characteristics at the population vector $K=(K_1, \ldots, K_R)$. Note that

$$\prod_{r=1}^{R} (K_r+1)$$

recursion steps have to be performed. This remark gives a first indication for the complexity of the MVA algorithm.

We shall describe a recursion step in the MVA algorithm and concentrate at the evaluation of the characteristics at the population vector $K = (K_1, \ldots, K_R)$. Afterwards the initialization of the recursion at $K=(0,...,0)$ is discussed.

A recursion step consists of three phases: 1. an evaluation of characteristics for closed chains, 2. an evaluation of marginal queuelength distributions, and 3. an evaluation of characteristics for open chains.

Phase 1 : Closed chain characteristics

In the first phase the three characteristics of closed chains are computed by means of the Relations (3.3.5), (3.3.6) and (3.3.7).

Step 1 : Expected residence times

Relation (3.3.5) yields for chain r, $r \in C$, at resource n, $n \in Q$,

$$S_{n,r}(K) = \sum_{k=1}^{M_n(K)} \frac{k w_{n,r}}{\mu_n(k)} p_n(k-1, K-e_r). \tag{3.4.1}$$

This evaluation may involve a summation of an infinite series. For special cases the evaluation reduces to a finite summation. As examples we mention the FCFS, PS and IS service disciplines as introduced in Subsection (3.3.5). In the next subsection we discuss the scheme for these special cases in more detail.

Step 2 : Throughputs

From Relation (3.3.6) it follows for $r \in C$ and $n \in Q$

$$\Lambda_{n,r}(K) = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(K)}. \tag{3.4.2}$$

Step 3 : Expected numbers of customers

From Relation (3.3.7) it follows for $r \in C$ and $n \in Q$

$$L_{n,r}(K) = \Lambda_{n,r}(K) S_{n,r}(K). \tag{3.4.3}$$

Phase 2 : Marginal queue length distributions

The second phase in the algorithm consists of the computation of the marginal queue length distributions by means of Relations (3.3.11) and (3.3.12).

The following auxiliary quantities are introduced for all $k = 1,...,M_n(K)$

$$a_n(k) = \sum_{i \in O} \frac{w_{n,i} \Lambda_{n,i}}{\mu_n(k)} , \tag{3.4.4}$$

$$A_n(k) = \prod_{l=1}^{k} a_n(l) , \tag{3.4.5}$$

$$b_n(k,K) = \sum_{r \in C} \frac{w_{n,r} \Lambda_{n,r}(K)}{\mu_n(k)} p_n(k-1, k-e_r) . \tag{3.4.6}$$

Furthermore, it is convenient to let $a_n(0) = A_n(0) = 1$. Observe that the values $a_n(k)$ are independent of the population vector and may be evaluated in advance, whereas the values $b_n(k,K)$ cannot be computed before phase 1 of the recursion step has been completed.

The probabilities $p_n(k,K)$ are the solution of, cf. (3.3.11) and (3.3.12),

$$p_n(k,K) = a_n(k) p_n(k-1,K) + b_n(k,K) , \tag{3.4.7}$$

for $k = 1,...,M_n(K)$, and

$$\sum_{k=0}^{M_n(K)} p_n(k,K) = 1 . \tag{3.4.8}$$

The solution of these equations is

$$p_n(k,K) = A_n(k) \left[p_n(0,K) + \sum_{l=1}^{k} \frac{b_n(l,K)}{A_n(l)} \right] , \tag{3.4.9}$$

where

$$p_n(0,K) = \frac{1 - \sum_{k=1}^{M_n(K)} A_n(k) \sum_{l=1}^{k} \frac{b_n(l,K)}{A_n(l)}}{\sum_{k=0}^{M_n(K)} A_n(k)} . \tag{3.4.10}$$

It is noted that the ergodicity condition (3.2.7) guarantees that the solution is a proper distribution. This is not a priori obvious from (3.4.10).
Furthermore, the evaluation of the marginal queue length distributions, involves a summation of an infinite series, unless the service rate function takes a special form.

Phase 3 : Characteristics of the open chains

The characteristics of the open chains may be evaluated in a straightforward way from (3.3.8), (3.3.9) and (3.3.10), i.e. for all open chains $i \in O$ we have at resource $n \in Q$

$$S_{n,i}(K) = \sum_{k=1}^{\infty} \frac{k w_{n,i}}{\mu_n(k)} p_n(k-1,K) \tag{3.4.11}$$

$$\Lambda_{n,i}(K) = \Lambda_{n,i} \tag{3.4.12}$$

$$L_{n,i}(K) = \Lambda_{n,i} S_{n,i}(K) \tag{3.4.13}$$

This completes the description of a recursion step in the MVA algorithm.

Initialization

The MVA algorithm starts with the computation of the characteristics of the open chains at population vector $\mathbf{0}=(0,\ldots,0)$, i.e. there are no closed chain customers in the system.

If a resource n is not visisted by customers of an open chain, i.e. $\Lambda_{n,i}=0$, for all $i \in O$, we initialize $p_n(0,\mathbf{0})=1$.
So, consider a resource that is visited by customers of open chains. The difference equation (3.4.9) is homogeneous and the solution is

$$p_n(k,0) = A_n(k) p_n(0,0)\,, \tag{3.4.14}$$

where

$$p_n(0,\mathbf{0}) = \frac{1}{\sum_{k=1}^{\infty} A_n(k)}\,. \tag{3.4.15}$$

Note that the marginal queue length distribution is proper, if the ergodicity condition (3.2.6) is satisfied. One may prove, cf. Reiser and Kobayashi [1975], that this condition guarantees the queueing network model to be ergodic for any population vector K. So, the mixed multichain queueing network model is ergodic, if an associated strictly open network is ergodic.

Once the marginal queue length distributions have been evaluated, one may evaluate the characteristics for the open chain customers by means of Relations (3.4.11) till (3.4.13).

The algorithm sketched above is the basis for an implementation of the Mean Value Analysis. However, it must be noted that for general service rate functions it is hard to design an efficient algorithm. Only for special functions the algorithm obtains a tractable form.

3.4.3. The MVA algorithm: fixed service rate functions

For the queueing network model introduced in Subsection 3.3.5 the MVA algorithm takes a simple form. The model is a frequently used tool in the analysis of a wide variety of queueing network systems. So, it is worthwhile to discuss its MVA algorithm in more detail.
It will be shown that the algorithm for the mixed multichain model reduces to an

adjusted model for a closed multichain queueing network model.

The MVA algorithm is a recursive scheme that runs through all vectors in the range of (0,...,0) to $(K_1, \ldots, K_R)$. A single recursion step, consists of two phases: 1. the evaluation of characteristics of closed chains, and 2. the evaluation of characteristics of open chains. The evaluation of the marginal queue length distribution is not needed.

First, we sketch the relations to be evaluated in a recursion step.

Phase 1 : Characteristics of the closed chains

The expected residence times, throughputs and expected numbers of customers of an closed chain $r \in C$ at the different resources may be evaluated from the Relations (3.3.45), (3.3.50), (3.3.40) and (3.3.41).
For the expected residence times we have at a resource n with a FCFS or PS service discipline

$$S_{n,r}(K) = \Big(\sum_{l \in O \cup C} L_{n,l}(K - e_r) + 1\Big) w_{n,r} \tag{3.4.16}$$

and at a resource n with an IS service discipline

$$S_{n,r}(K) = w_{n,r} \ . \tag{3.4.17}$$

Furthermore the throughputs are related by

$$\Lambda_{n,r}(K) = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(K)} \tag{3.4.18}$$

and the expected numbers of customers by

$$L_{n,r}(K) = \Lambda_{n,r}(K) S_{n,r}(K) \ . \tag{3.4.19}$$

Phase 2 : Characteristics of the open chains

The expected residence times, throughputs and expected numbers of customers of an open chain $i \in O$ may be evaluated from the Relations (3.3.44), (3.3.50), (3.3.42) and (3.3.43).
For the expected residence times we have at a resource n with a FCFS or PS service discipline

$$S_{n,i}(K) = \Big(\sum_{l \in O \cup C} L_{n,l}(K) + 1\Big) w_{n,i} \tag{3.4.20}$$

and at a resource n with an IS service discipline

$$S_{n,i}(K) = w_{n,i} \ . \tag{3.3.21}$$

The throughputs and expected numbers of customers are related by

$$\Lambda_{n,i}(K) = \Lambda_{n,i} \tag{3.4.22}$$

and

$$L_{n,i}(K) = \Lambda_{n,i} S_{n,i}(K) . \tag{3.4.23}$$

The Relations (3.4.20) through (3.4.23) form a linear system of equations for the characteristics of the open chain customers. This system has a simple solution. From Relations (3.4.20) and (3.4.23) it follows that

$$S_{n,i}(K) = (\sum_{l \in C} L_{n,l}(K) + \sum_{j \in O} \Lambda_{n,j} S_{n,j}(K) + 1) w_{n,i} . \tag{3.4.24}$$

This relation can be rewritten as

$$\frac{S_{n,i}(K)}{w_{n,i}} = \sum_{l \in C} L_{n,l}(K) + \sum_{j \in O} \Lambda_{n,j} w_{n,j} \frac{S_{n,j}(K)}{w_{n,j}} + 1 . \tag{3.4.25}$$

The right hand side does not depend on the index i and so

$$S_{n,i}(K) = \frac{\sum_{l \in C} L_{n,l}(K) + 1}{1 - \sum_{j \in O} \Lambda_{n,j} w_{n,j}} w_{n,i} . \tag{3.4.26}$$

Using Relation (3.4.26) instead of (3.4.20), we have a strictly recursive scheme for the evaluation of system characteristics for closed and open chains.

Initialization is by

$$S_{n,r}(0) = \Lambda_{n,r}(0) = L_{n,r}(0) = 0 \tag{3.4.27}$$

for all $r \in C$ and $n = 1,...,N$.

The second part of this subsection is devoted to a further simplification of the algorithm. Relation (3.4.26) suggests that only the evaluation of the characteristics of the closed customer chains is of importance. From (3.4.26) and (3.4.23) we obtain for all open chains $i \in O$ at resources n with a FCFS or PS service discipline

$$L_{n,i}(K) = \frac{\sum_{l \in C} L_{n,l}(K) + 1}{1 - \sum_{j \in O} \Lambda_{n,j} w_{n,j}} \Lambda_{n,i} w_{n,i} . \tag{3.4.28}$$

Inserting this relation in (3.4.16), we obtain after some algebra

$$S_{n,r}(K) = (\sum_{l \in C} L_{n,l}(K - e_r) + 1) \frac{w_{n,r}}{1 - \sum_{j \in O} \Lambda_{n,j} w_{n,j}} . \tag{3.4.29}$$

The characteristics of the closed customer chains thus can be evaluated from a closed multichain queueing network model with adjusted expected service demands. The characteristics of the open chains may be evaluated afterwards, cf. (3.4.26)

3.4.4. An implementation of the MVA algorithm

In this subsection we review the implementation of the recursion that forms the basis of the MVA algorithm. We discuss a straightforward implementation of the recursion which is based on a lexicographic enumeration of the set of population vectors. For a

more detailed analysis of implementations of the MVA algorithm we refer to Zahorjan and Wong [1981] and Balbo and Bruell [1980]. The overlay algorithm proposed in Zahorjan and Wong [1981] is, in priniple, more efficient than our implementation, but up till now an efficient implementation of the algorithm has not been found.

We shall describe the implementation for a given population vector $K=(K_1,\ldots,K_R)$.

The multi-dimensional recursion runs through all the integer valued vectors $(k_1,\ldots,k_R)$ in the range of $(0,\ldots,0)$ till $(K_1,\ldots,K_R)$. We denote the set of vectors in this range by

$$T = \{k=(k_1,\ldots,k_R) \mid k_r \in \{0,\ldots,K_r\},\ r=1,\ldots,R\}. \tag{3.4.30}$$

The crucial point in implementing the recursion is the design of a so-called feasible enumeration of the set T. An enumeration is feasible if $\boldsymbol{k}$ is preceded in the enumeration by $\boldsymbol{k}-\boldsymbol{e}_r$ for all $r=1,\ldots,R$.
We propose a transformation of the multi-dimensional recursion into a one-dimensional enumeration. This is done by a lexicographic ordering of the vectors in the set T.

Let the integers X_r, $r=1,\ldots,R+1$, be defined as

$$X_r = \prod_{l=1}^{r-1} (K_l + 1) \tag{3.4.31}$$

Note that the number of elements in the set T equals X_{R+1}.

Next the map $\phi : T \rightarrow \{0,\ldots,X_{R+1}-1\}$ is, for all $\boldsymbol{k} \in T$, defined by

$$\phi(\boldsymbol{k}) = \sum_{r=1}^{R} k_r X_r \tag{3.4.32}$$

where $\boldsymbol{k} \in T$.

The map ϕ will be used to construct a feasible enumeration of the set T. The enumeration is based on a linear enumeration of the set $\{0,\ldots,X_{R+1}-1\}$. The following two lemmas provide the basis of the enumeration. The first lemma shows that the map ϕ is one-to one and, as an aside, provides the inverse map ϕ^{-1} of ϕ. The second lemma shows the enumeration to be feasible.

Lemma 3.1

The map $\phi : T \rightarrow \{0,\ldots,X_{R+1}-1\}$ is one-to-one.

Proof
For $m \in \{0,\ldots,X_{R+1}-1\}$ define the vector $k(m)=(k_1(m),\ldots,k_R(m))$ by the following recursion in $r=R,R-1,\ldots,1$

$$k_r(m) = \left\lfloor \frac{m - \sum_{l=r+1}^{R} k_l(m) X_l}{X_r} \right\rfloor, \tag{3.4.33}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller or equal to x and where the empty sum is

defined to be zero.
It is readily verified that $\boldsymbol{k}(m) \in T$ and that $\phi(\boldsymbol{k}(m)) = m$.
That the map ϕ is one-to-one follows from the observation that the number of elements in the set T equals the number of elements in the set $\{0,\ldots,X_{R+1}-1\}$.

∎

Lemma 3.2

For all $\boldsymbol{k} \in T$ with $k_r > 0$ for some r the following equation holds

$$\phi(\boldsymbol{k}) = \phi(\boldsymbol{k} - \boldsymbol{e}_r) + X_r \,. \tag{3.4.34}$$

Proof

$$\phi(\boldsymbol{k}) = \sum_{l=1}^{R} k_l X_l - X_r + X_r = \phi(\boldsymbol{k} - \boldsymbol{e}_r) + X_r \,.$$

∎

The enumeration algorithm has the following structure:

```
Algorithm.

for m=0 step 1 until X_{R+1}-1 do
begin

    k := φ^{-1}(m);

    evaluate the characteristics at the population vector k
end.
```

It is easily verified that the number of recursion steps in this enumeration equals

$$\prod_{r=1}^{R} (K_r + 1) \,.$$

This implies that the computational complexity of the corresponding MVA algorithm grows exponentially with the number of closed customer chains. It is obvious that the number of recursion steps will form a handicap in the evaluation of systems with a larger number of closed customer chains.
Observe, that the complexity of a recursion step is highly dependent on the structure of the queueing network model and that it is possible that infinite series have to be summed, cf. Relations (3.4.1), (3.4.9), (3.4.10) and (3.4.11).

With respect to the storage requirements we remark that for a simple implementation of the algorithm in each recursion step information has to be stored and kept in memory. This implies that the required storage facilities grow exponentially in the number of closed customer chains as well. The overlay algorithm, cf. Zahorjan and Wong [1981], is more efficient than the standard MVA algorithms in that the storage requirements are less.

4. CLOSED MULTICHAIN QUEUEING NETWORK MODELS

4.1 Introduction

This chapter is concerned with the approximate evaluation of performance measures in closed multichain queueing network models with many closed customer chains. The exact evaluation may be performed by the MVA-algorithm, but computational complexity and the storage requirements restrict the use of the algorithm to models with relatively few closed customer chains. The complexity is largely caused by the fact that the number of recursion steps grows exponentially with the number of closed customer chains.
In this chapter we develop approximation methods for queueing network models with many closed customer chains. The analysis is restricted to models with first-come first-served, processor sharing and infinite server service disciplines. It is assumed that the service rates are queue length independent and normalized to unity. The main purpose is to show that the arguments sketched in Chapter 2 lead to the development of efficient and accurate approximation methods in a natural way.

The approximation methods are based on the mean value analysis approach and properties of the production and demand structure. The methods may be classified in three approaches.

The first approach is to avoid the recursion which causes the computational problems, and to concentrate on the evaluation of the performance measures at the desired population vector. Typical representatives are the widely appraised Schweitzer method, introduced in Schweitzer [1979], and its refinements suggested in Chandy and Neuse [1982].
In a wider perspective the approach may be viewed as a top-down analysis. Whereas the MVA-algorithm starts at the bottom vector (0,...,0) and ends at the desired population vector $(K_1, \ldots, K_R)$, one might consider the reverse way. In Eager and Sevcik [1983] and [1984] this idea has been worked out in a so-called Performance Bound Hierarchy method. We introduce a related concept : the depth improvement.

The second approach concerns a decomposition of the demand structure. Instead of the complicated network with many closed customer chains a simpler network model is evaluated for each customer chain. The mutual influence of the chains may be approximated for instance by an adjustment of service demands or service rates, cf. for instance Reiser [1979:1] and Reiser and Lavenberg [1980] or an adjustment of the mean value analysis procedure, cf. Van Doremalen [1984:1].
We discuss approximation methods along these lines. Apart from a depth improvement one might consider a partial decomposition to improve on the approximations. The idea is to split the set of closed customer chains in disjoint subsets that are treated, separately, as smaller closed multichain queueing network models.

The third approach involves an aggregation of the demand structure. We present a recursive aggregation-disaggregation method based on a complete aggregation of all closed

customer chains into one chain. A natural refinement is formed by a partial aggregation.

The chapter is organized as follows. In Section 4.2 we recapitulate the MVA-algorithm and the necessary notations. The Sections 4.3, 4.4 and 4.5 cover the three approaches. In Section 4.6 efficiency and accuracy of the methods and their refinements are discussed by some typical models stemming from the analysis of communication networks and computer systems.

4.2. Model and mean value analysis algorithm

For lucidity of presentation a multichain queueing network model is introduced and the corresponding mean value analysis algorithm is recapitulated.

The service discipline at the N resources is first-come first-served or processor sharing. The service rate of the single service unit at resource $n = 1,...,N$ is fixed and normalized to unity.
The K_r customers of the closed customer chain r, $r = 1,...,R$, proceed through the network in accordance with a Markov routing given by an irreducible stochastic matrix P_r. The visiting ratios $f_{n,r}$ of customers of chain r at resource n are the unique stationary probabilities of the discrete Markov chains associated with these routing matrices. The service demands at the resources are independent random variables with a mean $w_{n,r}$ for customers of chain r at resource n. At each first-come first-served resource the demands must be exponentially distributed with the same mean for all customers.

The following steady state characteristics play a role:

$S_{n,r}(K)$ expected residence time of a customer of chain r at resource n,

$\Lambda_{n,r}(K)$ throughput of customers of chain r at resource n, and

$L_{n,r}(K)$ expected number of customers of chain r at resource n,

where K emphasizes the dependence on the population vector.

In Chapter 3 the mean value analysis algorithm has been described for the computation of these characteristics. It constitutes the following three relations:

$$S_{n,r}(K) = \left(\sum_{i=1}^{R} L_{n,i}(K-e_r) + 1\right)w_{n,r} , \tag{4.2.1}$$

$$\Lambda_{n,r}(K) = \frac{f_{n,r}K_r}{\sum_{m=1}^{N} f_{m,r}S_{m,r}(K)} , \tag{4.2.2}$$

$$L_{n,r}(K) = \Lambda_{n,r}(K)S_{n,r}(K) . \tag{4.2.3}$$

The recursion is initialized by $L_{n,r}(0)=0$, for all $n = 1,...,N$ and $r = 1,...,R$, and runs through all vectors in the range of $(0,...,0)$ through $(K_1, \ldots, K_R)$. It is easily verified that

$$\prod_{r=1}^{R}(K_r+1)$$

recursion steps have to be executed. Only for small values of $R, K_1, \ldots, K_R$ an exact evaluation of the algorithm will be possible. The design of approximation methods is the way out that will be analysed in the sequel of this chapter.

4.3. Removal of the recursion

4.3.1. Introduction

This section deals with a first approach for the approximation of behavioural characteristics in closed multichain queueing network models. The computational complexity of the MVA-algorithm is largely caused by the recursion. A reduction of the number of recursion steps seems a good starting point for the development of approximation methods.

In its most extreme form this will lead to the formulation of a set of approximating mean value relations at the desired population vector. We introduce basic methods which have their roots in the techniques presented in Schweitzer [1979] and Chow [1983].
In a more general framework the approach can be formulated as a top-down analysis. Whereas the exact MVA-algorithm runs through all the population vectors in the range of (0,...,0) till $(K_1, \ldots, K_R)$, in the approximation methods only the last steps in this recursion are performed. This idea will be worked out in detail. The emphasis is on the formulation of a so-called first order depth improvement.
Most of the approximation methods in this line may be formulated as successive approximation methods for the determination of a fixed point of a non-linear operator. Techniques to ascertain the existence and uniqueness of the solutions and the convergence of the iteration schemes are discussed. One of the new results is an alternative proof of the convergence of the Schweitzer method in case the model comprises only one closed customer chain.

4.3.2. The basic method

The recursion appears in the expected residence time relation only. A formal way to remove the recursion from the scheme is by the introduction of a set of auxiliary quantities, cf. also Reiser and Lavenberg [1980], Chandy and Neuse [1982] and Chow [1983].

Defining $\epsilon_{n,i,r}(K)$ by

$$\epsilon_{n,i,r} = \frac{L_{n,i}(K)-L_{n,i}(K-e_r)}{L_{n,i}(K)}, \tag{4.3.1}$$

we may rewrite Relation (4.2.1) as

$$S_{n,r}(K) = \left[\sum_{i=1}^{R}[1-\epsilon_{n,i,r}(K)]L_{n,i}(K)+1\right]w_{n,r}. \tag{4.3.2}$$

So, the quantities $\epsilon_{n,i,r}(K)$ are the relative difference between the expected number of customers of chain i at resource n for the population vectors K and $K-e_r$. The last step in the recursive MVA-algorithm has been rewritten in terms of the population vector K only.

However, for an exact evaluation of the performance measures at vector K the quantities $\epsilon_{n,i,r}(K)$ have to be computed by means of the complete original MVA-algorithm. To bypass this complex computation it is suggested to use approximations for these quantities. Before doing so, the last step in the MVA-algorithm is rewritten leaving out the argument K.
The Relations (4.2.1), (4.2.2) and (4.2.3) become

$$S_{n,r} = (\sum_{i=1}^{R} (1-\epsilon_{n,i,r})L_{n,i} + 1)w_{n,r} , \tag{4.3.3}$$

$$\Lambda_{n,r} = \frac{f_{n,r}K_r}{\sum_{m=1}^{N} f_{m,r}S_{m,r}} , \tag{4.3.4}$$

$$L_{n,r} = \Lambda_{n,r}S_{n,r} . \tag{4.3.5}$$

A first guess for $\epsilon_{n,i,r}$ might be

$$\epsilon_{n,i,r} = 0 , \tag{4.3.6}$$

for all $n=1,...,N$ and $i,r=1,...,R$. The rational is that for larger population sizes the expected numbers of customers at the distinct resources will not change much with the removal or addition of a customer. In Chow [1983] this approximation has been worked out in detail.

An interpretation in the line of the mean value approach gives some insight in where the method may fail. The idea is based on the assumption that a customer at a jump moment observes the system as if in equilibrium. For smaller population sizes this argument will yield bad approximations.

It is rather straightforward to amend for one obvious mistake that is introduced by the approximation idea. The arrival theorem states that a jumping customer observes the system as if in equilibrium with one customer of its own chain removed. If it is assumed that the distribution of the customers over the resources does not change with the addition or removal of one customer, the following approximation, introduced in Schweitzer [1979], makes sense

$$\epsilon_{n,i,r} = \frac{\delta(i,r)}{K_r} , \tag{4.3.7}$$

where $\delta(i,r)$ is the Kronecker δ.

The behavioural characteristics may thus be approximated by a solution of a set of non-linear equations, namely

$$S_{n,r} = (\sum_{i=1}^{R} L_{n,i} - \frac{L_{n,r}}{K_r} + 1)w_{n,r} , \tag{4.3.8}$$

$$\Lambda_{n,r} = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}} , \tag{4.3.9}$$

$$L_{n,r} = \Lambda_{n,r} S_{n,r} . \tag{4.3.10}$$

A standard way to solve this system of equations is by successive approximation. With initial values for $L_{n,r}$ the scheme defined by (4.3.8), (4.3.9) and (4.3.10) is repeatedly evaluated until convergence has been established. Though it is, in principle, possible that the method diverges, the method converges in all experimental situations. This remarkable result has been backed up with relatively few theoretical results. In Subsection 4.3.4 we return to this subject.

4.3.3. Depth improvement

The Schweitzer method has attracted a lot of attention as it is highly efficient and fairly accurate. However, if a higher accuracy is required and one is willing to spend extra computation time and storage facilities, a straightforward depth improvement may be considered.

The basic idea is to evaluate approximations for the behavioural characteristics at the population vectors $K - e_r$, $r = 1,...,R$, and then to evaluate the last step in the recursive MVA-algorithm exactly. We shall refer to this improvement as a first order depth improvement. It is a special case of a range of improvements which we have called depth improvements.

The MVA-algorithm runs through all vectors in the range of (0,...,0) through $(K_1, \ldots, K_R)$. With V we denote the set of all R-dimensional integer valued vectors in this range, i.e.

$$V := \{ \boldsymbol{k} = (k_1, \ldots, k_R) \mid k_r \in \{0,...,K_r\} , r = 1,...,R \} . \tag{4.3.11}$$

The MVA-algorithm is recursive in the number of customers of the distinct chains. So, it makes sense to introduce the number T as

$$T = \sum_{r=1}^{R} K_r \tag{4.3.12}$$

and the sets $V(k)$, $k = 0,1,...,T$, as

$$V(k) = \{ \boldsymbol{k} \in V \mid \sum_{r=1}^{R} k_r = k \} . \tag{4.3.13}$$

Observe that the sets $V(k)$ are disjoint and their union is the set V. These sets form the basis for an efficient implementation of the MVA-algorithm as well, cf. Zahorjan and Wong [1981]. To evaluate the characteristics at the vectors in the set $V(k)$ the

behavioural characteristics at the population vectors in the set $V(k-1)$ are needed only.

We now introduce the kth-order depth improvement. First, approximations for the behavioural characteristics at the population vectors in the set $V(T-k)$ are evaluated. Afterwards, an evaluation of the original MVA-algorithm is performed for the population vectors in the sets $V(T-k+1)$ through $V(T)$.
The depth improvement is a generally applicable method that may be used in virtually all methods discussed in this chapter, and it yields very satisfactory results. A more elaborated first order depth improvement of the Schweitzer method has been introduced in Chandy and Neuse [1982].
Another interesting use of higher order depth improvements is the Performance Bound Hierarchy method introduced in Eager and Sevcik [1983] and [1984]. The method generates upper and lower bounds on certain performance characteristics by an appropriate choice of initial sets of characteristics.

4.3.4. Notes on convergence, existence and uniqueness

In this subsection some remarks are made with respect to the existence, and uniqueness of the solution of the non-linear equations (4.3.3) through (4.3.5) and the convergence of the corresponding successive approximation scheme for the special case of a single closed customer chain.

In the recent literature this problem has received some attention, cf. for instance Reiser and Lavenberg [1980], De Souza e Silva, Muntz and Lavenberg [1983] and Chow [1983].
In Chow [1983] the approximation suggested by (4.3.6) is analysed in detail. It is proved that the solution of the corresponding set of non-linear equations is unique.
We will analyse the general case. The existence of a fixed point may be proved by an application of Brouwer's fixed point theorem. For the uniqueness of the fixed point and the convergence of the successive approximation method only marginal results are known. For the queueing network model with exactly one closed customer chain, we give a new proof for the convergence of the successive approximations to a unique and positive solution. The proof is based on the theory of non-negative matrices.

Before we proceed with the discussion, a formulation of Brouwer's fixed point theorem is introduced.

Theorem 4.1

Let F be a continuous map of a compact and convex subset $D \in R^n$ into itself. Then the equation $F(x) = x$ has at least one solution in D.

Proof : cf. Ortega and Rheinboldt [1970:161-162]. ∎

We shall rewrite the set of non-linear system of equations (4.3.3), (4.3.4) and (4.3.5) in such a way that it takes the form of an eigenvalue-eigenvector problem. It is assumed throughout the sequel of this subsection that the values of the quantities $\epsilon_{n,i,r}$ satisfy

$0 \leqslant \epsilon_{n,i,r} \leqslant 1$, for all indices n, i, r.
The equations can be rewritten as a set of equations for the values $L_{n,r}$ in the following way. For $n = 1,\ldots,N$ and $r = 1,\ldots,R$,

$$L_{n,r} = \frac{(\sum_{i=1}^{R}[1 - \epsilon_{n,i,r}]L_{n,i} + 1)f_{n,r}w_{n,r}}{\sum_{m=1}^{N}(\sum_{i=1}^{R}[1 - \epsilon_{m,i,r}]L_{m,i} + 1)f_{m,r}w_{m,r}}. \tag{4.3.14}$$

Defining the auxiliary quantities $x_{n,r}$ and $\phi_{n,r}$ as

$$x_{n,r} = \frac{L_{n,r}}{K_r}$$

and

$$\phi_{n,r} = \frac{f_{n,r}w_{n,r}}{\sum_{m=1}^{N} f_{m,r}w_{m,r}},$$

a set of non-linear equations for the quantities $x_{n,r}$, $n = 1,\ldots,N$ and $r = 1,\ldots,R$, is obtained:

$$x_{n,r} = \frac{(\sum_{i=1}^{R}K_i[1 - \epsilon_{n,i,r}]x_{n,i} + 1)\phi_{n,r}}{\sum_{m=1}^{N}(\sum_{i=1}^{R}K_i[1 - \epsilon_{m,i,r}]x_{m,i} + 1)\phi_{m,r}}. \tag{4.3.15}$$

Note that $x_{n,r}$ may be interpreted as an approximation for the long run fraction of the number of customers of chain r being at resource n.

Let us introduce the row-vector notation

$$x = (x_1, \ldots, x_R)$$

for the quantities $x_{n,r}$, where, for $r = 1,\ldots,R$,

$$x_r = (x_{1,r}, \ldots, x_{N,r}).$$

Furthermore, the set $D \subset R^{NR}$ is defined as

$$D = \{ y \in R^{NR} \mid 0 \leqslant y_{n,r} \leqslant 1 \text{ and } \sum_{m=1}^{N} y_{m,r} = 1 \}. \tag{4.3.16}$$

Note that D is a compact and convex subset of R^{NR}. Now, the map $F: R^{NR} \to R^{NR}$ is defined by its components $F_{n,r}$ as

$$F_{n,r}(y) = \frac{(\sum_{i=1}^{R}K_i[1 - \epsilon_{n,i,r}]y_{n,i} + 1)\phi_{n,r}}{\sum_{m=1}^{N}(\sum_{i=1}^{R}K_i[1 - \epsilon_{m,i,r}]y_{m,i} + 1)\phi_{m,r}}. \tag{4.3.17}$$

It is easily verified that the map F is continuous on D and maps D into itself. Invoking the fixed point theorem we conclude that there exists at least one $x^* \in D$ such that $F(x^*) = x^*$. This guarantees the existence of at least one solution of the original problem in the feasible region D.

Of course it would have been nice if we would also have been able to prove the uniqueness of the solution and the convergence of the successive approximation method. Regrettably, we have not been able to do so for the general case, but we have succeeded in constructing a new proof for uniqueness and convergence of the successive approximations in the case of a single closed customer chain.

As, for $r = 1,\ldots,R$,

$$\sum_{m=1}^{N} x_{m,r} = 1 ,$$

we can rewrite (4.3.15) as,

$$x_{n,r} = \frac{\sum_{i=1}^{R} \left[K_i [1 - \epsilon_{n,i,r}] x_{n,i} + \frac{1}{R} \sum_{j=1}^{N} x_{j,i} \right] \phi_{n,r}}{\sum_{m=1}^{N} \sum_{i=1}^{R} \left[K_i [1 - \epsilon_{m,i,r}] x_{m,i} + \frac{1}{R} \sum_{j=1}^{N} x_{j,i} \right] \phi_{m,r}} \tag{4.3.18}$$

Defining, for all $i,r = 1,\ldots,R$, matrices $A^{i,r} \in R^{N \times N}$ with elements $A^{i,r}_{m,n}$ by

$$A^{i,r}_{m,n} = \begin{cases} \left[K_i [1 - \epsilon_{n,i,r}] + \frac{1}{R} \right] \phi_{n,r} & , \quad n = m \\ \frac{1}{R} \phi_{n,r} & , \quad n \neq m \end{cases} \tag{4.3.19}$$

we may rewrite (4.3.18), for each $r = 1,\ldots,R$ in vector-matrix notation as

$$x_r = \frac{\sum_{i=1}^{R} x_i A^{i,r}}{\sum_{i=1}^{R} x_i A^{i,r} e} , \tag{4.3.20}$$

where e is the column vector of ones with a suitable dimension.

For $R = 1$ we may drop the subscripts i and r and it may be assumed that ϕ_n is strictly positive for all $n = 1,\ldots,N$. This leaves us with the problem of finding a positive solution of,

$$x = \frac{xA}{xAe} . \tag{4.3.21}$$

Observe that the matrix A has strictly positive elements. As a consequence of the Perron-Frobenius theory for non-negative matrices, cf. Seneta [1973] or Berman and Plemmons [1979], we may conclude that equation (4.3.21) has a unique and strictly positive

solution and that the successive approximation method, defined by Relation (4.3.15), converges geometrically. The asymptotic rate of convergence equals $|\sigma_1|/|\sigma_2|$, where σ_1 and σ_2 are the in absolute value largest and second largest eigenvalue of the matrix A respectively.

4.4. Decomposition of the demand structure

4.4.1. Introduction

This section deals with a second approach for the approximation of behavioural characteristics in closed multichain queueing network models. The complexity of the mean value analysis algorithm is largely due to the fact that the demand structure comprises a large number of closed customer chains. It seems therefore natural to use an aggregation or decomposition of the demand structure. In the next section we concentrate on aggregation methods, whereas here the emphasis is on decomposition methods.

The basic idea behind a decomposition of the demand structure is to construct R queueing network models with a single chain and an adjusted set of parameters or an adjusted mean value analysis procedure. Subsection 4.4.2 treats methods which are based on an adjustment of the parameter sets. In Subsection 4.4.3 an adjustment of the mean value analysis algorithm is considered.

Again, it is possible to design improvements of the standard methods. The depth improvement introduced in the preceding section is such an improvement. We shall design a partial decomposition method as well in Subsection 4.4.4.

4.4.2. Adjustment of the parameters

The first line is based on adjusting the parameters of the R parallel single chain queueing network models. To the customers of a particular chain it appears as if the service units at the resources are slowed down because of the work that these have to perform for the customers of the other chains. This may be accounted for by a service demand or service rate adjustment.

The methods are split in a decomposition and a composition step and proceed as follows.

The decomposition step starts with the analysis of a particular chain r. Assume that an initial set of adjusted service demands $\tilde{w}_{n,r}$ has been given for all $n=1,\ldots,N$. A one-dimensional MVA-algorithm has to be evaluated.
For $k=1,\ldots,K_r$ the relations at $n=1,\ldots,N$ are given by

$$S_{n,r}(k) = (L_{n,r}(k-1)+1)\tilde{w}_{n,r}\ , \tag{4.4.1}$$

$$\Lambda_{n,r}(k) = \frac{f_{n,r}k}{\sum_{m=1}^{N} f_{m,r}S_{m,r}(k)}\ , \tag{4.4.2}$$

$$L_{n,r}(k) = \Lambda_{n,r}(k) S_{n,r}(k) . \tag{4.4.3}$$

The recursion, as usual, is started with $L_{n,r}(0)=0$ for all $n=1,...,N$. For each chain r this computation yields new approximations.

The composition step comprises the construction of a new set of service demands $\tilde{w}_{n,r}$. We shall discuss two alternatives.

The first alternative has been suggested in in Reiser [1979:2] and in a somewhat different context in Reiser [1976] and Sevcik [1978].
In the decomposition step approximations $\Lambda_{n,i}(K_i)$ have been computed for the throughput of customers of chain i at resource n. So, the fraction of time that the service unit at resource n is serving customers of chain i is approximated by $\Lambda_{n,i}(K_i)w_{n,i}$.
A plausible adjustment of the service demand of a customers of chain r is given by

$$\tilde{w}_{n,r} = \frac{w_{n,r}}{1 - \sum_{i \neq r} \Lambda_{n,i}(K_i) w_{n,i}} . \tag{4.4.4}$$

The denominator at the right hand side may be interpreted as the fraction of time the service unit at resource n is free for the handling of customers of chain r.

Regrettably, the method presented above cannot be recommended as it is possible that the composition step yields negative or even undefined values $\tilde{w}_{n,r}$. This may be amended by a simple trick. Note that the problem arises if at a certain resource n,

$$\sum_{i=1}^{R} \Lambda_{n,i}(K_i) w_{n,i} \geqslant 1 .$$

This can be remedied by replacing (4.4.4) with

$$\tilde{w}_{n,r} = \frac{w_{n,r}}{1 - \dfrac{\sum_{i \neq r} \Lambda_{n,i}(K_i) w_{n,i}}{\sum_{i=1}^{R} \Lambda_{n,i}(K_i) w_{n,i}}} . \tag{4.4.5}$$

The decomposition and composition step suggest an iterative procedure. Using (4.4.5) instead of (4.4.4) yields in each iteration step feasible approximations for the characteristics. As in the preceding section the existence of a fixed point can be proved applying Brouwers fixed point theorem. The uniqueness of the solution and convergence of the iteration scheme have not been established.

The second composition step is based on an idea introduced in Reiser and Lavenberg [1980]. Though in this paper no intuitive argument is given, the intuitive background of the method seems to be formed by the interpretation of the processor sharing service discipline, cf. Subsection 3.3.5.

In the decomposition step approximations $L_{n,i}(K_i)$ have been computed for the expected number of customers of chain i at resource n. Now

$$\frac{L_{n,r}(K_r)}{\sum_{i=1}^{R} L_{n,i}(K_i)} \tag{4.4.6}$$

is an approximation for the fraction of the number of customers at resource n being of chain r. If resource n would have had a processor sharing service discipline this fraction could have been used as an approximation for the fraction of the total service rate dedicated to customers of chain r. An plausible adjustment of the service demand of a customer of chain r at resource n therefore seems to be

$$\tilde{w}_{n,r} = \left[\frac{L_{n,r}(K_r)}{\sum_{i=1}^{R} L_{n,i}(K_i)} \right]^{-1} w_{n,r} \ . \tag{4.4.7}$$

Again, we have formulated the basis for an iterative decomposition/composition method. In each iteration step feasible performance characteristics are constructed. Following the line in Subsection 4.3.4, it is not difficult to establish the existence of a solution of the implicitly defined set of non-linear equations for these performance characteristics.

In Reiser [1979:1] and Reiser and Lavenberg [1980] another line is followed. There the basic approximation idea is a removal of the recursion as suggested by the Relations (4.3.1) through (4.3.5). The approximate values for the quantities $\epsilon_{n,i,r}$ are obtained from the above described decomposition steps in the following way.

After an evaluation of the decomposition step a two-stage composition step is performed. First, the quantities $\epsilon_{n,i,r}$ are approximated by,

$$\epsilon_{n,i,r} = \begin{cases} \dfrac{L_{n,i}(K_i) - L_{n,i}(K_i - 1)}{L_{n,i}(K_i)} & , \ i = r \\ 0 & , \ i \neq r \end{cases} \tag{4.4.8}$$

and (4.3.3), (4.3.4) and (4.3.5) are evaluated.
Then, new values for $\tilde{w}_{n,r}$ are computed from the new approximations for the behavioural characteristics. The adjustments are based on the (4.4.4) and (4.4.7).

For the discussion of the numerical results in Section 4.6 we have chosen for the original two approximation methods. The extra amount of work that has to be done by in evaluating (4.3.3) through (4.3.5) makes the more complicated methods numerically less attractive, whereas numerical experiments have shown that the accuracy is in the same order.

4.4.3. Adjustment of the mean value analysis procedure

The second approach is based on an adjustment of the MVA-algorithm and, more in particular, on an adjustment of the expected residence time relation (4.2.1).

Consider a particular chain r. To obtain approximations for the characteristics of this particular chain we perform a decomposition step based on the following MVA-algorithm for a network with one chain.
For $k = 1,2,...,K_r$ compute at all resources $n = 1,...,N$

$$S_{n,r}(k) = (L_{n,r}(k-1) + A_{n,r}(k-1) + 1)w_{n,r} , \tag{4.4.9}$$

$$\Lambda_{n,r}(k) = \frac{f_{n,r}k}{\sum_{m=1}^{N} f_{m,r}S_{m,r}(k)} , \tag{4.4.10}$$

$$L_{n,r}(k) = \Lambda_{n,r}(k)S_{n,r}(k) , \tag{4.4.11}$$

where the recursion starts with $L_{n,r}(0)=0$ for all $n = 1,...,N$. The term $A_{n,r}(k-1)$ in (4.4.9) represents the expected number of customers of other chains a customer of chain r sees in front of it on arrival at resource n if the population vector is $\boldsymbol{K} - (K_r - k)\boldsymbol{e}_r$.

It is not difficult to verify from (4.2.1), (4.2.2) and (4.2.3) that the adjusted single chain MVA-algorithm yields exact results, if one puts, for all $k = 1, \ldots, K_r$,

$$A_{n,r}(k-1) = \sum_{i \neq r} L_{n,i}(\boldsymbol{K} - (K_r - (k-1))\boldsymbol{e}_r) . \tag{4.4.12}$$

However, the computation of the values $A_{n,r}(k)$ implies the evaluation of the complete exact MVA-algorithm and that is just what we want to avoid.

We suggest, therefore, to use an approximation for the values $A_{n,r}(k)$. The idea is to assume that a customer of chain r sees at a jump moment the situation for the other chains as if in equilibrium.

For $A_{n,r}(k)$ we therefore put for $k = 1,2,...,K_r$,

$$A_{n,r}(k) = \sum_{i \neq r} L_{n,i}(K_i) . \tag{4.4.13}$$

Note that this step defines a composition step in an iterative decomposition/composition method and that the values $A_{n,r}(k)$ do not depend on k any longer. The decomposition step is started with an initial set of approximations for $A_{n,r}(k)$. Afterwards, in a composition step, the new values of $A_{n,r}(k)$ are evaluated.

The iterative scheme again defines a successive approximation method for the determination of a solution of a set of non-linear equations. Using the same arguments as in Subsection 4.3.4 one may prove the existence of a solution. We have not been able to proof the convergence of the method or the uniqueness of the solution, but the method converges in all practical situations.

Numerical experiments have also shown that the method converges faster than the ones presented in the preceding subsection, that the iteration shows a less wild behaviour and that the approximations are less often really bad. In fact, one might say that it is more robust. A possible explanation is provided by the observation that the approximations for $A_{n,r}(k)$ satisfy an important conservation property: the sum of all these values equals the sum of all customers in the other chains, as it should be.

4.4.4. Partial decomposition

In the preceding subsections we have discussed methods based on a complete decomposition of the demand structure. All chains have been analysed separately in a decomposition step. A natural relaxation of this method is a partial decomposition which is based on a partitioning of the set of closed customer chains in subsets.
Though numerical experiments have indicated that the depth improvement as suggested in Subsection 4.3.4 yields in a more efficient way equally accurate approximations, the partial decomposition method is presented as a natural relaxation step in the design of approximation methods. However, in the discussion of the numerical results in Section 4.6 the method is not reviewed.

As before let $C = \{1,...,R\}$ be the set of closed customer chains. A partitioning of C in I disjoint subsets, $C_1, \ldots, C_I$ say, provides a basis for a partial decomposition. Assume that the chains $r_{i-1}+1, \ldots, r_i$ are in the set C_i for $i = 1,...,I$, where $r_0 = 0$ by definition. This can always be achieved be a renumbering of the chains. With each subset C_i, $i = 1,...,I$ we associate a population vector $\boldsymbol{B}_i$

$$B_i = (K_{r_{i-1}+1},...,K_{r_i}) . \tag{4.4.14}$$

The partial decomposition method proceeds as follows.

The decomposition step is initialized by a set of approximations for the expected numbers of customers of the distinct chains at the distinct resources. For $r \in C_i$ these are denoted as $L_{n,r}(\boldsymbol{B}_i)$. For each subset C_i of chains a MVA-algorithm is evaluated running through all the vectors $\boldsymbol{b}_i$ in the range of the null-vector and $\boldsymbol{B}_i$. The scheme is given by the following three mean value relations that are very similar to (4.4.9), (4.4.10) and (4.4.11),

$$S_{n,r}(\boldsymbol{b}_i) = \Big(\sum_{l \in C_i} L_{n,l}(\boldsymbol{b}_i - \boldsymbol{e}_r) + \sum_{j \neq i} \sum_{l \in C_j} L_{n,l}(\boldsymbol{B}_j) + 1 \Big) w_{n,r} , \tag{4.4.15}$$

$$\Lambda_{n,r}(\boldsymbol{b}_i) = \frac{f_{n,r} k_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(\boldsymbol{b}_i)} , \tag{4.4.16}$$

$$L_{n,r}(\boldsymbol{b}_i) = \Lambda_{n,r}(\boldsymbol{b}_i) S_{n,r}(\boldsymbol{b}_i) . \tag{4.4.17}$$

The recursion is started with $L_{n,r}(\boldsymbol{0}) = 0$ for all $r \in C_i$ and $n = 1,...,N$.

The composition step is simple since new approximations for the expected numbers of customers have been computed in the decomposition step.

An important measure for the numerical complexity of the method is the required number of recursion steps per iteration step. It is not difficult to verify that, for a given partitioning $C_1, \ldots, C_I$, this number equals

$$\sum_{i=1}^{I} \prod_{r \in C_i} (K_r + 1) .$$

In applications one will only consider a partial decomposition with a few subsets. For example, to study the characteristics of a particular chain one might consider a partitioning in two subsets, where one set contains this particular chain and the other set the remaining chains. Furthermore, it should be noted that $I = 1$ corresponds with the exact MVA-algorithm and $I = R$ with a complete decomposition.

4.5. Aggregation of the demand structure

4.5.1. Introduction

This section concerns a third approach towards the approximation of large and closed multichain queueing network models. The method discussed is based on a recursive aggregation-disaggregation of the demand structure.
In Subsection 4.5.2 a complete aggregation is considered, whereas 4.5.3 deals with an improvement by a partial aggregation.
The methods are strictly recursive. This implies that existence, uniqueness and convergence problems do not have to be answered. Furthermore, the resulting algorithms are finite and it is possible to give an a priori estimation for the amount of computation time and the size of storage facilities needed. These features make the aggregation methods to a very promising and interesting new tool for the approximate analysis of large scale queueing network systems. In Chapter 6 we shall see that an application of the methods in the more complex context of a network system with priority queues yields very good result as well.

4.5.2. Basic aggregation method

In this subsection we introduce an approximation method based on a recursive aggregation-disaggregation of the demand structure. Starting point of the analysis is the MVA-algorithm as given by (4.2.1), (4.2.2) and (4.2.3).

It is suggested to perform a complete aggregation of the set of closed customer chains into one closed customer chain. As a consequence the adjusted MVA-algorithm for the aggregation method runs through the integers $0,\ldots,T$, where

$$T = \sum_{r=1}^{R} K_r . \tag{4.5.1}$$

In each step of this single chain recursion both an aggregation and a disaggregation step are performed.
The disaggregation uses an adjustment of (4.2.2) and (4.2.3). The basic idea is that the customers which are lumped together in a single chain in an aggregation step, may be distributed over the original chains in a disaggregation step. Let α_r, $r = 1,\ldots,R$, be the fraction of the total number of customers belonging to chain r, i.e.

$$\alpha_r = \frac{K_r}{T} . \tag{4.5.2}$$

We now introduce the recursive aggregation-disaggregation method in more detail. The basis is a recursion in the integer values $k = 1,\ldots,T$, where each recursion step comprises an aggregation and a disaggregation step.

A recursion step starts with an aggregation step. If it is assumed that at a jump moment a customer observes the system as if in equilibrium with one customer removed, the expected total number of customers that a customer sees present on arrival at a resource n may be approximated by

$$L_n(k-1) = \sum_{r=1}^{R} L_{n,r}(k-1) . \tag{4.5.3}$$

The disaggregation step comprises the evaluation of the following three adjusted mean value relations that are very similar to (4.2.1), (4.2.2) and (4.2.3)

$$S_{n,r}(k) = (L_n(k-1) + 1)w_{n,r} , \tag{4.5.4}$$

$$\Lambda_{n,r}(k) = \frac{\alpha_r f_{n,r} k}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(k)} , \tag{4.5.5}$$

$$L_{n,r}(k) = \Lambda_{n,r}(k) S_{n,r}(k) . \tag{4.5.6}$$

The recursion may be started with $L_n(0) = 0$, for all $n = 1,\ldots,N$.

The algorithm is easy to implement and has the considerable advantage over most existing approximation methods that it is non-iterative. It is evident that the method may be improved by the simple trick of a depth improvement. In the next subsection we discuss another obvious improvement: a partial aggregation method.

4.5.3. Partial aggregation

In this subsection we discuss a refinement of the global aggregation method. As in the partial decomposition method it seems a good idea to use a partitioning of the set of chains. Each of the created subsets forms the basis for an aggregate closed customer chain.

A partitioning of C in I subsets, $C_1, \ldots, C_I$ say, is the basis of the partial aggregation method. The sets C_i are disjoint and their union is C. With the subset C_i we associate a

population B_i, where

$$B_i = \sum_{r \in C_i} K_r \ . \tag{4.5.7}$$

This defines a population vector $B = (B_1, \ldots, B_I)$ for the aggregated model. We introduce for all $r \in C_i$ the quantity α_r as the fraction of the number of customers in the aggregate C_i which is of chain r, i.e.

$$\alpha_r = \frac{K_r}{B_i} \ . \tag{4.5.8}$$

The approximating MVA-algorithm again is an adjusted mean value analysis scheme involving an I-dimensional recursion on the population vector B.

The recursion is defined by the following four relations at an arbitrary population vector b in the range of (0,...,0) through $(B_1, \ldots, B_I)$.
For all $r \in C_i$, $i = 1,\ldots,I$ and $n = 1,\ldots,N$ evaluate

$$S_{n,r}(b) = (L_n(b - e_i) + 1) w_{n,r} \ , \tag{4.5.9}$$

$$\Lambda_{n,r}(b) = \frac{\alpha_r b_i f_{n,r}}{\sum_{m=1}^{N} f_{m,r} S_{m,r}(b)} \ , \tag{4.5.10}$$

$$L_{n,r}(b) = \Lambda_{n,r}(b) S_{n,r}(b) \ , \tag{4.5.11}$$

$$L_n(b) = \sum_{r=1}^{R} \Lambda_{n,r}(b) L_{n,r}(b) \ . \tag{4.5.12}$$

The recursion runs through all the vectors in the range of (0,...,0) through $(B_1, \ldots, B_I)$ and, consequently,

$$\prod_{i=1}^{I} (B_i + 1)$$

recursion steps are to be evaluated. The amount of work per recursion step is very similar to that in the original MVA-algorithm. As a consequence the total amount of work to be done will be much less than for the original MVA-algorithm if the value of I is not too large.

We have introduced the partial aggregation method. The problem that remains, is a good choice of a partitioning. The accuracy of the method will tend to be better for a more detailed partitioning, but a refined partitioning will cost more computation time and larger storage facilities than a simple one. One has to look for a partitioning in the range of $I = R$ corresponding with the exact MVA-algorithm, and $I = 1$ corresponding with the global aggregation, which yields the right balance between desired accuracy and acceptable computational costs.

The number of recursion steps gives a good indication for the computational complexity and may be used to give an a priori estimation of the efficiency of a given partitioning.
The accuracy of the method for a given partitioning is hard to measure as no bounds are presently known for the resulting procedures. Here, numerical experiments and an intuitive insight have to be combined in order to be able to come up with some general guide lines.
It has appeared that for general purposes the first order depth improvement is a reliable tool for improving the accuracy of the method. For more detailed problems the use of a partial aggregation may be considered. We present some examples in Section 4.6.

4.6. Numerical examples

4.6.1. Introduction

For none of the proposed approximation methods it is clear on forehand whether it will yield an acceptable accuracy, as no error-bounds have been derived. To compare the methods we, therefore, present three examples stemming from the analysis of information processing systems.

The first example concerns a model of a communication network with a window-flow control protocol. The second example concerns a model of a time-sharing system comprising a set of terminals, a central processor unit and a set of I/O-devices. The third example concerns a closed central server model comprising multiple central processor units sharing a set of I/O-devices.

We have restricted the analysis to separable queueing network models. This allows for a comparison of the approximate and the exact results and the conclusions are not confused by the extra complication of a violation of the separability conditions.
Such an extra complication is introduced in Chapter 6 in which we study large scale closed multichain queueing network models in which the separability conditions are violated by the introduction of priority schedules.

4.6.2. A communication network with window flow control

In this subsection we study closed multichain queueing network models of communication networks with a window flow control protocol. For more details on such systems and a validation of some of the assumptions we refer to Reiser [1979:2] and Lam and Wong [1982].

A communication network is designed for the transportation of messages between so-called hosts, e.g. main frame computers, single terminals and large data bases. It is unattractive to couple every pair of hosts by a dedicated channel. So, one shall consider the installation of a communication network comprising a set of switches and a network of channels connecting these switches. The switches are highly specialized software driven

devices which control the traffic of the messages through the network. A message enters the network at a source-switch and proceeds through the network from switch to switch via the intermediate channels until it reaches a destination-switch. It is assumed that the path followed by a message is determined by the source and destination host. A message is buffered at a switch until the next channel on its route becomes free.

If every message is allowed for to enter the system, the network will get clogged up and extreme long response times are the unsatisfactory consequence. On the other hand, if only a single message is allowed for to be in the system, the capacity of the network will be ill-used and the response times will therefore be undesirable long. Apparently, it is an interesting problem to design a protocol forming the in-between of these extreme schedules.

A typical admittance schedule achieving this goal is the window flow control protocol. The basic idea is that for each source-destination pair a maximum number of messages, the window, is allowed for to be in the network. If this maximum number is reached, newly arriving messages are buffered in a so-called source buffer until a place in the network becomes available. This protocol is relatively easy to implement as the source switch receives acknowledgements for all (un)successfully transmitted messages.

We introduce a closed multichain queueing network model of a communication network with window flow control.

The model comprises N channels which are modelled as resources with a single service unit operating at a fixed service rate normalized to unity. It is common practice to model the delays at the switches as integral parts of the service demands at the channels. Though the service discipline at the channels, in principle, is first-come first-served, it is convenient to assume a processor sharing service discipline, since this allows for non-exponential service demand distributions with a different mean for the different types of messages, cf. Reiser [1979:2]. Another way to bypass this problem might be the use of the approximation method suggested in Subsection 2.5.2.

The network provides service to a set of hosts which communicate with the network as medium. A source-destination pair forms a virtual channel. It is assumed that the route of a message through the communication network is fixed for a given source-destination pair. Thus a virtual channel is given by a source host, a set of (physical) channels and a destination host.
We discern R virtual channels. The route of a message of the virtual channel r, $r=1,...,R$, is implicitly defined by a set of visiting ratios $f_{n,r}$, $n=1,...,N$. Here, $f_{n,r}=1$ if the message visits channel r and $f_{n,r}=0$ otherwise. The service demands of the messages of virtual channel r at channel n are stochastically independent random variables with mean $w_{n,r}$.

For each virtual channel r we introduce the window size K_r which indicates the maximum number of messages that is allowed for to be in the network.

We consider two ways to model the arrival processes of messages to the communication network.

The first way uses the introduction of a source queue, cf. Reiser [1979:2]. The arrival process is modelled as a resource with a first-come first-served resource with a single service unit and a fixed service rate normalized to unity. The service demands at the source queue of virtual channel r are stochastically independent and exponentially distributed with mean w_r. Note that the number of messages in the network for each virtual channel r may vary from 0 to K_r.

The second way uses the assumption that the system operates under heavy traffic, cf. Lam and Wong [1982]. This implies that at the moment a message leaves the network at the destination switch, instantaneously a new message enters the network at the source switch. Note that in this way the number of messages in the network equals the window size K_r for every virtual channel r at every instant.
In fact, this modelling may be viewed as a special case of the first one by setting the service demands at the source queues to 0.

Observe that both ways of modelling lead to separable and closed multichain queueing network models. The resources are the source queues and the channels. The customer chains are the virtual channels. Such models tend to be very large as the number of channels and virtual channels are in the order of several dozens. The exact evaluation of such models is prohibited by the computational complexity of the MVA algorithm and the use of approximation methods has to be considered.

In the sequel of this subsection we present two numerical examples that are based on the two ways of modelling the arrival processes. The discussion concentrates at the evaluation of three important performance characteristics: the expected response time of a message, the throughput of messages, and the utilization of the channels. Apart from an exact evaluation with an implementation of the MVA algorithm the characteristics have been pictured for seven approximation methods:

SW	the Schweitzer method described in Subsection 4.3.2,
SW-DI	the Schweitzer method with a first order depth improvement,
DSDA	the decomposition method with a service demand adjustment suggested by Relation (4.4.4),
DSRA	the decomposition method with a service rate adjustment suggested by Relation (4.4.7),
DMVA	the decomposition method based on a mean value analysis extension suggested in Subsection 4.4.3,
AG	the global aggregation method suggested in Subsection 4.5.2,

AG-DI the global aggregation method with a first order depth improvement.

The stop criterion for the iterative methods is a five-decimal precision in the maximal relative difference between two successive approximations of the throughputs. This choice seems to be quite arbitrary, but numerical experiments have shown the results to be only marginal influenced by the choice of a higher precision.

The model parameters for the two examples are pictured in Tables 4.1 and 4.5. The first example uses the concept of source queues, whereas the second example studies the system under heavy traffic.

The Tables 4.2, 4.3 and 4.4 concern the first example and give the exact and approximate results for the expected response times and throughputs of the messages of the virtual channels and the utilizations of the physical channels. We have pictured the processing times in seconds in the last row of Table 4.2 (CPT).
The Tables 4.6 and 4.7 picture expected response times, utilizations and processing times for the second example.

At first glance, the figures show that all methods yield quite good approximations. A first tentative conclusion which has been strengthened by other experiments, is that the relatively simple Schweitzer method or the global aggregation method have to be recommended. If these methods are used with a first order depth improvement, the results are within a few percent of the exact values.
The decomposition methods perform equally well. Numerical experiments have shown that the third method has to be advocated. This iterative mean value analysis extension converges fast and the convergence does not show a wild behaviour in the first few iteration steps. This in contrast with the first two methods.
A disadvantage of the iterative methods is that it is difficult to find a good stopping criterion and that it is not clear beforehand how many iteration steps are needed.
In this respect, it should be remarked that the method of Schweitzer and the decomposition method with the mean value extension are rather robust. Already after a couple of iteration steps these methods yield a reasonably acceptable accuracy.

The presentation of the examples in the 4.6.3 and 4.6.4 support these observations.

virtual channel	window size	service demand source queue	service demand channels			
			1	2	3	4
1	6	2	2	2	2	0
2	8	2	0	0.5	0.5	0.5
3	4	3	4	4	0	0
4	8	5	1	0	0	1

Table 4.1 : A communication network model (example 1).

virtual channel	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	28.98	29.19	28.62	29.22	28.32	29.80	29.26	27.92
2	5.73	5.55	5.64	5.60	5.37	5.91	5.36	5.42
3	47.17	48.74	46.72	48.52	47.06	49.72	49.01	45.23
4	7.66	7.34	7.45	7.58	6.74	7.51	7.22	7.32
CPT	6.25	0.02	0.06	0.50	0.60	0.15	0.02	0.08

Table 4.2 : Expected response times of the messages (example 1).

virtual channel	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	.188	.187	.190	.186	.191	.184	.185	.194
2	.491	.471	.482	.478	.453	.493	.440	.473
3	.079	.076	.079	.077	.079	.075	.076	.082
4	.200	.195	.198	.198	.188	.200	.180	.185

Table 4.3 : Throughput of the virtual channels (example 1).

physical channel	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	.890	.874	.896	.876	.884	.867	.853	.910
2	.936	.915	.939	.917	.923	.913	.893	.952
3	.621	.610	.621	.611	.608	.613	.590	.625
4	.445	.430	.439	.437	.414	.446	.400	.431

Table 4.4 : Utilizations of the physical channels (example 1).

virtual channel	window size	service demand channels											
		1	2	3	4	5	6	7	8	9	10	11	12
1	2	1	0	0	0	0	1	1	0	0	1	0	0
2	2	1	1	0	0	1	0	0	0	0	0	1	0
3	2	0	1	1	0	0	1	1	1	1	0	0	0
4	2	0	0	1	0	1	0	0	0	1	1	1	0
5	2	1	0	1	1	1	0	0	0	0	0	0	0
6	2	0	0	0	1	0	0	0	1	1	1	1	1
7	2	1	0	0	0	0	1	1	1	1	0	0	0
8	2	0	0	0	0	0	0	0	1	1	1	1	1

Table 4.5 : A communication network model (example 2).

virtual channel	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	8.89	9.43	8.87	9.41	9.43	9.43	9.97	8.89
2	8.82	9.34	8.82	9.31	9.34	9.34	9.86	8.81
3	12.79	13.15	12.67	13.18	13.13	13.15	13.69	12.77
4	11.35	11.84	11.37	11.84	11.83	11.84	12.27	11.36
5	8.14	8.58	8.16	8.55	8.60	8.58	9.17	8.14
6	13.27	13.58	13.30	13.58	13.58	13.58	13.99	13.28
7	12.18	12.77	12.23	12.74	12.84	12.77	13.01	12.13
8	11.70	12.05	11.75	12.05	12.04	12.05	12.46	11.70
CPT	32.00	0.05	0.30	0.25	0.25	0.15	0.05	0.30

Table 4.6 : Expected response times of messages (example 2).

physical channel	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	.861	.816	.861	.818	.815	.816	.775	.862
2	.629	.599	.630	.600	.599	.599	.567	.629
3	.578	.554	.579	.555	.554	.554	.527	.578
4	.396	.380	.396	.381	.380	.380	.361	.396
5	.403	.383	.403	.384	.383	.383	.366	.403
6	.545	.521	.547	.521	.520	.520	.500	.546
7	.545	.521	.547	.521	.520	.520	.500	.546
8	.642	.622	.642	.622	.622	.622	.603	.643
9	.818	.791	.818	.791	.791	.791	.766	.819
10	.723	.694	.722	.695	.695	.694	.667	.723
11	.725	.696	.723	.697	.697	.696	.669	.725
12	.322	.313	.321	.313	.314	.313	.303	.322

Table 4.7 : Utilizations of the physical channels (example 2).

4.6.3. A time-sharing system

A basic model in the analysis of computer systems is the central server model, cf. for instance Sauer and Chandy [1981] and Lavenberg and Sauer [1983]. The central part of such a model comprises a set of central processor units and I/O-devices: the computer system or central processor. The computer system executes jobs which are generated by different types of customers, for instance edit and compile jobs from terminals, large batch jobs and data-base enquiries from remote sources.

In this subsection we present an example of a central server model comprising one central processor unit, a set of three I/O-devices and three groups of terminals, see Figure 4.1.

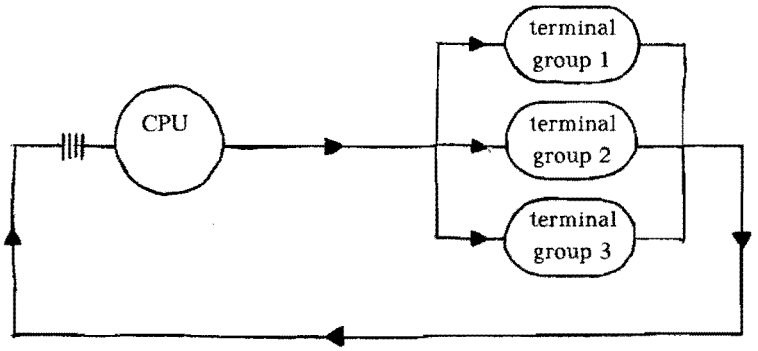

Figure 4.1 : A central server model with terminal groups.

The model assumptions are as follows. The users at the 20 terminals of group 1 have exponentially distributed think times with a mean of 10 seconds and generate jobs that on the average comprise 20 visits to the CPU, 15 to I/O-1 and 4 to I/O-2. The users at the 10 terminals of group 2 have exponentially distributed think times with a mean of 20 seconds and generate jobs that on the average comprise 40 visits to the CPU, 14 to I/O-1 and 25 to I/O-2. The users at the 10 terminals of group 3 have exponentially distributed think times with a mean of 60 seconds and generate jobs that on the average comprise 200 visits to the CPU, 20 to I/O-1, 40 to I/O-2 and 139 to I/O-3.
The CPU and the I/O-devices are modelled as first-in first-out resources with a fixed service rate that is normalized to unity. The mean service demands per visit to the CPU, I/O-1, I/O-2 and I/O-3 are 10 msec, 20 msec, 20 msec and 30 msec respectively. The service demands are stochastically independent and exponentially distributed.

The model is a separable closed multichain queueing network model with three customer chains. In the Tables 4.8. 4.9 and 4.10 we have pictured the results of an exact evaluation and seven approximation methods, cf. the discussion in the preceding subsection.

The numerical results agree with the observations that we have made in the preceding subsection.

terminal group	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	1.69	1.81	1.75	1.74	1.84	1.77	1.71	1.70
2	3.11	3.29	3.21	3.22	3.34	3.26	3.12	3.11
3	17.30	18.43	17.81	17.79	19.31	17.95	18.27	17.63
CPT	5.50	0.02	0.05	0.25	0.25	0.25	0.05	0.15

Table 4.8 : Expected response times of the jobs at the computer.

terminal group	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
1	1.711	1.694	1.702	1.703	1.690	1.699	1.708	1.710
2	0.433	0.429	0.431	0.431	0.428	0.430	0.433	0.433
3	0.129	0.127	0.129	0.129	0.126	0.128	0.128	0.129

Table 4.9 : Throughput of jobs at the computer.

resource	exact	removal recursion		decomposition			aggregation	
		SW	SW-DI	DSDA	DSRA	DMVA	AG	AG-DI
CPU	.774	.766	.770	.770	.761	.768	.770	.773
I/O-1	.686	.679	.683	.683	.677	.682	.684	.686
I/O-2	.457	.452	.454	.454	.450	.454	.455	.456
I/O-3	.539	.532	.536	.526	.526	.535	.533	.537

Table 4.10 : Utilization of the computer system resources.

4.6.4. A closed central server model

The third example considers a closed central server model of a computer system comprising multiple central processor units or CPU's and a set of commonly shared I/O devices or I/O's. Such models are of importance when studying computer systems with a buffer or memory queue and a multi-programming service discipline by means of a decomposition technique, cf. Courtois [1977] and Hine, Mitrani and Tsur [1979].

We consider one example. The numerical results are representative for this type of models. As an aside we discuss tentative conclusions with respect to the use of a partial aggregation method for this type of queueing network models.

The system comprises three CPU's and nine I/O's. The service discipline at the CPU's is processor sharing and at the I/O's first-come first-served. The service rates at the resources are fixed and normalized to unity.

There are three types of jobs. The jobs of a given type are distributed over the CPU's. The numbers of jobs per type and per CPU are pictured in Table 4.11. The expected service demands of the customers at the CPU's depend on the type and CPU number and are pictured in Table 4.11 also.
After a visit to the CPU a visit to an I/O device follows. The probabilities of visiting a given I/O depend on the type number only and are pictured in Table 4.12. The service demands at an I/O device are exponentially distributed with an expectation that is independent of the type or CPU number of a given customer. The expectations are pictured in Table 4.12.

The resulting model is a separable closed multichain queueing network model with twelve resources and nine closed customer chains.
In the Tables 4.13 and 4.14 we have pictured the numerical results of a set of approximation methods. In Table 4.13 the throughputs of the different types of customers at the distinct CPU's are pictured. In Table 4.14 the utilizations of the CPU's are pictured and as an aside the processing times of the methods are given in seconds.

The numerical results support the discussion in the preceding subsections.

We have added two new approximations, PA-1 and PA-2, which are based on a partial aggregation of the set of customer chains. As we have three types of jobs which are divided over the three CPU's two obvious aggregations offer themselves immediately.
The first one partitions the set of nine chains in three subsets corresponding with the three types. So, the aggregate i, $i = 1,2,3$, comprises all customers of type i. The results are pictured under the name of PA-1.
The second one partitions the set of chains in three subsets corresponding with the three CPU's. So, the aggregate i, $i = 1,2,3$, comprises all those customers which are dedicated to CPU i. The results are pictured under the name of PA-2.

It should be observed that the second partitioning yields better results than the first one. This may be explained by a more general observation: an aggregation yields good results if the chains which are lumped together, show some structural resemblance. For the approximate analysis of large closed multichain queueing network models it has appeared that the routing is of paramount influence on the accuracy. If chains which visit approximately the same set of resources, are lumped together, the approximate results of the partial aggregation method tend to be good

Further research in this direction is necessary to validate these first tentative conclusions.

	type 1			type 2			type 2		
cpu	1	2	3	1	2	3	1	2	3
population	3	1	1	2	2	1	1	2	3
demand	10	15	15	20	25	30	30	40	50

Table 4.11 : Populations and service demands at the CPU's.

	visiting probability per I/O			service
	type 1	type 2	type 3	demand
1	.25	.15	.00	30
2	.25	.15	.00	30
3	.25	.15	.05	30
4	.05	.10	.05	50
5	.05	.10	.15	50
6	.05	.10	.15	60
7	.06	.05	.20	60
8	.04	.05	.20	80
9	.00	.15	.20	80

Table 4.12 : Visiting probabilities and service demands at I/O's.

CPU	type 1			type 2			type 3		
	1	2	3	1	2	3	1	2	3
EXACT	32.03	9.47	9.05	13.39	12.24	5.67	4.72	8.34	10.82
SW	31.58	9.18	8.73	13.03	11.86	5.42	4.55	8.04	10.34
SW-DI	32.27	9.52	9.05	13.45	12.23	5.63	4.75	8.36	10.75
DSDA	31.66	9.20	8.79	13.06	11.92	5.39	4.54	8.08	10.50
DSRA	31.30	8.82	8.38	12.78	11.36	5.01	4.48	7.71	9.71
DMVA	31.64	9.19	8.78	13.04	11.90	5.37	4.53	8.06	10.48
AG	30.48	8.66	8.25	12.70	11.32	5.08	4.41	11.32	9.63
AG-DI	32.26	9.35	8.86	13.54	12.14	5.53	4.82	8.33	10.61
PAG-1	31.49	8.75	8.34	12.90	11.50	5.07	4.40	7.69	9.96
PAG-2	31.55	9.25	8.90	13.06	12.11	5.59	4.51	8.18	10.68

Table 4.13 : Throughputs at the CPU's per type.

	utilizations			processing
	CPU-1	CPU-2	CPU-3	time in sec
EXACT	.730	.782	.847	12.50
SW	.713	.756	.811	0.05
SW-DI	.734	.783	.840	0.50
DSDA	.714	.759	.819	0.60
DSRA	.703	.724	.762	0.75
DMVA	.713	.758	.817	0.15
AG	.691	.717	.758	0.02
AG-DI	.738	.777	.829	0.30
PAG-1	.705	.726	.775	0.50
PAG-2	.712	.769	.835	0.50

Table 4.14 : Utilizations of the CPU's and processing times of the algorithmic procedures.

5. QUEUEING NETWORK MODELS WITH TWO PHASE SERVERS

5.1. Introduction

This chapter treats the approximate analysis of queueing network models with a special type of two phase servers. At some of the resources the service demand splits in two phases, where the first one is preparatory and may be executed in the absence of the customer. However, the preparatory phase may be executed in advance for one customer only. So only the first customer in a busy period may find its first phase already executed. Such resources do not satisfy the separability conditions and consequently the resulting queueing network model is non-separable. The analysis is of interest as such two-phase servers are a natural model for some phenomena in queueing network analysis. In fact, the model is a special case of a larger class of important queueing network models. It represents a system where the first customer of a busy period has a deviating service demand distribution, cf. Welch [1964].

In Section 1.4 it has been observed that separable queueing networks, essentially, may be applied in two ways for the construction of approximation methods. The first approach is to approximate general queueing network models by separabale queueing network models, cf. the discussion in Section 2.4. The second approach is to design approximating adjustments and extensions of the exact evaluation procedures for separable queueing network models. Especially the MVA algorithm with its attractive interpretation forms a good starting point for such methods, cf. the discussion in Section 2.5.

In this chapter three approximation methods are presented.

The first method is based on the first approach and is an application of the iterative aggregation-disaggregation method which has been introduced in Subsection 2.4.2.
The second method is based on a straightforward mean value analysis extension and is related to the approximation method discussed in Subsection 2.5.2.
The third method combines the ideas of the two approaches. In an iterative procedure an approximating separable queueing network model is constructed. The formulation of a new set of parameters is based on the MVA approach. One of the more interesting parts in the analysis is the formulation of a set of conditions for which the iterative procedure converges. As a side result a monotonicity property of throughputs in a closed single chain queueing network model is established.

The chapter is organized as follows. In Section 5.2 the queueing network model is introduced and the mean value analysis of the corresponding separable queueing network is presented. The analysis of a single resource in such a queueing network model by means of a parametric analysis type of argument is discussed as well. The relation with the theory described in Chandy, Herzog and Woo [1975:1] is demonstrated.
The Sections 5.3 till 5.5 are dedicated to the three approximation methods. In Section 5.6 some numerical results are discussed and a few conclusions are drawn.

5.2. Model, mean value analysis and parametric analysis

In this section the queueing network model is introduced. Mean value analysis and parametric analysis of the basic separable queueing network model are reviewed.

The system comprises N resources with a service rate function μ_n, a dedication function ϕ_n and an admittance function γ_n at resource n, $n=1,...,N$. There are K customers in the network which belong to a single chain. The relative visiting ratios are given by f_n, $n=1,...,N$. The service demand at a resource n is exponentially distributed with expectation w_n. This description defines a separable queueing network with a single customer chain, if the separability conditions are satisfied, cf. Section 3.2 for more details.

However, a subset $A \subset \{1,...,N\}$ of servers operates differently. At a resource $n \in A$ a single service unit operates at a constant service rate which is normalized to unity. The service demand splits up in two independent and exponentially distributed phases with means v_n and w_n, respectively. The customers are served in order of arrival and the first phase has to be executed before the second one. However, if there are no customers at the resource the service unit may execute the first preparatory phase of the first customer in the next busy period. The second phase may only be executed when the customer is present.

The approximation methods are based on the mean value analysis approach. The parametric analysis of such models will play a role in the analysis of the methods.

First, the MVA algorithm is recapitulated for this single chain queueing network model with queue length dependent service rates. For details we refer to Chapter 3.
The algorithm starts with a relation for the expected residence times at resource n when K customers are in the system

$$S_n(K) = \sum_{k=1}^{K} \frac{k w_n}{\mu_n(k)} p_n(k-1,K-1) . \tag{5.2.1}$$

If the service rate is constant and normalized to unity, this relation may be simplified. At a first-come first-served or processor sharing queue we obtain

$$S_n(K) = L_n(K-1)w_n + w_n , \tag{5.2.2}$$

where $L_n(K-1)$ denotes the expected number of customers at resource n when $K-1$ customers are in the system. At an infinite server queue this yields

$$S_n(K) = w_n . \tag{5.2.3}$$

The second relation couples the throughput at resource n with the expected residence times

$$\Lambda_n(K) = \frac{f_n K}{\sum_{m=1}^{N} f_m S_m(K)} . \tag{5.2.4}$$

The third relation couples the marginal queue length distributions for the systems with K and $K-1$ customers. With $p_n(k,K)$ being the probability that k customers are at resource n when there are K customers in the system the relation is, for $k=1,\ldots,K$, given by

$$p_n(k,K) = \frac{w_n \Lambda_n(K)}{\mu_n(k)} p_n(k-1,K-1) \tag{5.2.5}$$

and, for $k=0$ by

$$p_n(0,K) = 1 - \sum_{k=1}^{K} p_n(k,K) . \tag{5.2.6}$$

The fourth relation couples the expected number of customers with the throughput and the expected residence time by

$$L_n(K) = \Lambda_n(K) S_n(K) . \tag{5.2.7}$$

These four relations comprise the recursive MVA algorithm.

The second part of this section considers a technique to evaluate in an efficient way the performance characteristics at a particular resource for varying parameters. It will be shown that the characteristics may be evaluated from an associated system comprising two resources: the given resource and a complementary resource describing the influence of the remaining part of the queueing network system. The analysis is based on the MVA approach and the results are very similar to the results in Chandy, Herzog and Woo [1975:1].

Let us concentrate on the evaluation of the performance characteristics at a given resource n. A complementary resource is associated with this particular resource. This resource has a first-come first-served service discipline and queue length dependent service rates $T_n(k)$, $k=1,\ldots,K$. The service demands are stochastically independent and exponentially distributed with unit mean. The K customers are alternately visiting the resource n and its complement.
With the marginal queue length probabilities at resource n being $p_n(k,K)$, for $k=0,\ldots,K$, the rates $T_n(k)$, $k=1,\ldots,K$, are defined by the flow rate relation

$$\frac{\mu_n(K-k+1)}{w_n} p_n(K-k+1,K) = T_n(k) p_n(K-k,K) , \tag{5.2.8}$$

which yields

$$T_n(k) = \frac{\mu_n(K-k+1)}{w_n} \frac{p_n(K-k+1,K)}{p_n(K-k,K)} . \tag{5.2.9}$$

It will be shown that the values $T_n(K)$ as defined by (5.2.9) do not depend on the model parameters at resource n. The proof uses a slightly adjusted version of Theorem 3.1.

The state of the single chain queueing system may be described by a vector $\boldsymbol{k}=(k_1,\ldots,k_N)$, where k_n denotes the number of customers present at resource n. The associated state space $S(K)$ is given as

$$S(K)=\{(k_1,\ldots,k_N) \mid 0\leqslant k_n \leqslant K \text{ and } \sum_{n=1}^{N} k_n=K\}. \tag{5.2.10}$$

The steady-state probabilities $p(\boldsymbol{k})$ of the system being in a state $\boldsymbol{k}\in S(K)$ are, cf. Theorem 3.1,

$$p(\boldsymbol{k})=\frac{1}{G(K)}\prod_{m=1}^{N}\frac{x_m^{k_m}}{\beta_m(k_m)}, \tag{5.2.11}$$

where

$$x_m = f_m w_m \tag{5.2.12}$$

and

$$\beta_m(k_m)=\prod_{k=1}^{k_m}\mu_m(k). \tag{5.2.13}$$

The normalization constant $G(K)$ is given by

$$G(K)=\sum_{\boldsymbol{k}\in S(K)}\prod_{m=1}^{N}\frac{x_m^{k_m}}{\beta_m(k_m)}. \tag{5.2.14}$$

Furthermore, we introduce the auxiliary quantities $G^{[n]}(K)$ as

$$G^{[n]}(K)=\sum_{\substack{\boldsymbol{k}\in S(K)\\ k_n=0}}\prod_{m=1}^{N}\frac{x_m^{k_m}}{\beta_m(k_m)}. \tag{5.2.15}$$

The following Lemma provides the prerequisites for the analysis.

Lemma 5.1

The following three relations hold for all $K\geqslant 0$, $n=1,\ldots,N$ and $i=0,\ldots,K$:

$$G(K)=\sum_{j=0}^{K}\frac{x_n^j}{\beta_n(j)}G^{[n]}(K-j), \tag{5.2.16}$$

$$p_n(i,k)=\frac{x_n^i}{\beta_n(i)}\frac{G^{[n]}(K-i)}{G(K)}, \tag{5.2.17}$$

$$\Lambda_n(K)=f_n\frac{G(K-1)}{G(K)}. \tag{5.2.18}$$

Proof :

$$G(K)=\sum_{\boldsymbol{k}\in S(K)}\prod_{m=1}^{N}\frac{x_m^{k_m}}{\beta_m(k_m)}$$

$$= \sum_{j=0}^{K} \sum_{\substack{k \in S(K) \\ k_n = j}} \prod_{m=1}^{N} \frac{x_m^{k_m}}{\beta_m(k_m)}$$

$$= \sum_{j=0}^{K} \frac{x_n^j}{\beta_n(j)} \sum_{\substack{k \in S(K-j) \\ k_n = 0}} \prod_{m=1}^{N} \frac{x_m^{k_m}}{\beta_m(k_m)}$$

$$= \sum_{j=0}^{K} \frac{x_n^j}{\beta_n(j)} G^{[n]}(K-j)$$

$$p_n(i,K) = \sum_{\substack{k \in S(K) \\ k_n = i}} \frac{1}{G(K)} \prod_{m=1}^{N} \frac{x_m^{k_m}}{\beta_m(k_m)}$$

$$= \frac{1}{G(K)} \frac{x_n^i}{\beta_n(i)} \sum_{\substack{k \in S(K-i) \\ k_n = 0}} \prod_{m=1}^{N} \frac{x_m^{k_m}}{\beta_m(k_m)}$$

$$= \frac{x_n^i}{\beta_n(i)} \frac{G^{[n]}(K)}{G(K)}$$

$$\Lambda_n(K) = \sum_{j=1}^{K} p_n(j,K) \frac{\mu_n(j)}{w_n}$$

$$= \sum_{j=1}^{K} \frac{x_n^j}{\beta_n(j)} \frac{G^{[n]}(K-j)}{G(K)} \frac{\mu_n(j)}{w_n}$$

$$= \frac{f_n}{G(K)} \sum_{j=1}^{K} \frac{x_n^{j-1}}{\beta_n(j-1)} G^{[n]}(K-1-(j-1))$$

$$= f_n \frac{G(K-1)}{G(K)}$$

∎

Thus, it follows from the definition of $T_n(k)$ and the results of Lemma 5.1 that, for $k = 1,\ldots,K$,

$$T_n(k) = f_n \frac{G^{[n]}(k-1)}{G^{[n]}(k)} . \tag{5.2.19}$$

This last relation corresponds with Relation (20) in Chandy, Herzog and Woo [1975:1]. It is easily seen that the values $G^{[n]}(k)$ do not depend on the parameters at resource n. As a consequence we can evaluate the performance characteristics at resource n for different sets of parameters by studying the system with the resource n and its

complement: the service rates of the complement need to be evaluated only once.

5.3. An iterative aggregation-disaggregation method

5.3.1. Introduction

In this section an iterative aggregation-disaggregation method is discussed. The line of argument follows the discussion in Subsection 2.4.2. The method is based on the recognition of two hierarchical levels. At the higher level a separable queueing network is analysed. At a lower level more detailed models are introduced for two phase server resources. An exact analysis of these models appears to be possible.
An analysis of the higher level yields a new set of parameters for the lower level model and vise versa. An iteration between these two levels yields approximations for some important system characateristics.

In Subsection 5.3.2 a detailed model for a two phase server is introduced and analysed. In Subsection 5.3.3 the iterative aggregation-disaggregation method is presented.

5.3.2. The two phase server model

The lower level detailed model of a two phase server is defined as a resource with a single service unit, queue length dependent Poisson instream rates, and a first-come first-served service discipline. The instream rates are $\lambda(k)$ for $k=0,\ldots,K-1$. The service demand splits into two independent and exponentially distributed phases, where the first phase is preparatory. The parameters of the exponential distributions are ν and μ respectively. The buffer has a capacity of K customers.

The state of the system is given by a tuple (k,f), where k, $k \in \{0,\ldots,K\}$ denotes the number of customers in the system and f, $f \in \{1,2\}$ the phase the service unit is executing. Note that $(0,1)$ corresponds with the situation that the service unit is executing the first phase of a customer that is not present and that $(0,2)$ describes the situation that the service unit is idle, but that the first phase of the service demand of the next customer to arrive has been completed.

The model describes an irreducible and time-homogeneous Markov process on a finite state space. The limiting probabilities $p(k,f)$ of the system being in some state (k,f) are the unique and strictly positive solution of the set of equilibrium equations and a normalization.

For $k=0$ these equations are

$$p(0,1)(\lambda(0)+\nu) = p(1,2)\mu \tag{5.3.1}$$

$$p(0,2)\lambda(0) = p(0,1)\nu\ , \tag{5.3.2}$$

for $k=1,\ldots,K-1$

$$p(k,1)(\lambda(k)+\nu) = p(k-1,1)\lambda(k-1) + p(k+1,2)\mu \tag{5.3.3}$$

$$p(k,2)(\lambda(k)+\mu) = p(k-1,2)\lambda(k-1) + p(k,1)\nu \tag{5.3.4}$$

and for $k = K$

$$p(K,1)\nu = p(K-1,1)\lambda(K-1) \tag{5.3.5}$$

$$p(K,2)\mu = p(K-1,2)\lambda(K-1) + p(K,1)\nu\ . \tag{5.3.6}$$

The normalization is given by

$$\sum_{k=0}^{K}\sum_{f=1}^{2} p(k,f) = 1\ . \tag{5.3.7}$$

This linear system may be solved recursively. Define $\pi(k,f)$ by

$$\pi(k,f) = \frac{p(k,f)}{p(0,1)}\ . \tag{5.3.8}$$

The values $\pi(k,f)$ satisfy a set of recursive relations. Then, for $k=0$ and $f=1$ the recursion is initialized by

$$\pi(0,1) = 1\ . \tag{5.3.9}$$

Matching the total flow-rate out of state (0,2) with the total flow-rate into state (0,2) yields

$$\pi(0,2) = \frac{\mu}{\lambda(0)}\ . \tag{5.3.10}$$

From the state transition diagram pictured in Figure 5.1 it follows that for $k = 1,2,\ldots,K$

$$\lambda(k-1)(\pi(k-1,1) + \pi(k-1,2)) = \nu\pi(k,2)\ , \tag{5.3.11}$$

or

$$\pi(k,2) = \frac{\lambda(k-1)}{\nu}(\pi(k-1,1) + \pi(k-1,2))\ . \tag{5.3.12}$$

Matching the total flow-rate out of state $(k,2)$ with the total flow-rate into state $(k,2)$ yields for $k = 1,\ldots,K$

$$(\nu + \lambda(k))\pi(k,2) = \lambda(k-1)\pi(k-1,2) + \mu\pi(k,1)\ , \tag{5.3.13}$$

where it is assumed that $\lambda(K)=0$. This yields for $k = 1,\ldots,K$

$$\pi(k,1) = \frac{\nu+\lambda(k)}{\mu}\pi(k,2) - \frac{\lambda(k-1)}{\mu}\pi(k-1,2)\ . \tag{5.3.14}$$

The Relations (5.3.9), (5.3.10), (5.3.12) and (5.3.14) define a recursive scheme for the evaluation of the quantities $\pi(k,f)$ for all $k=0,\ldots,K$ and $f=1,2$.
The limiting probabilities $p(k,f)$ are easily obtained as

$$p(k,f) = \frac{\pi(k,f)}{\sum_{i=0}^{K}(\pi(i,1)+\pi(i,2))}\ . \tag{5.3.15}$$

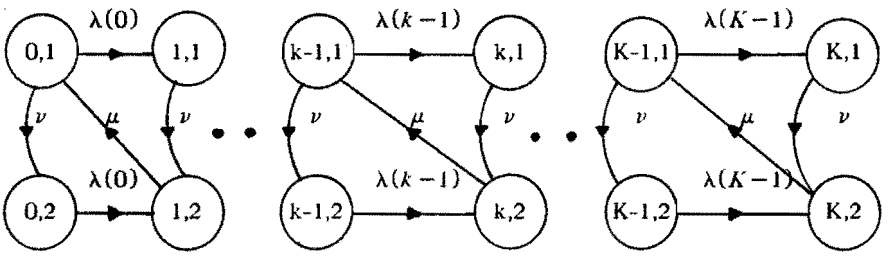

Figure 5.1 : State-transition diagram of the two phase server model.

This completes the exact analysis of a lower level model of a two phase server resource.

5.3.3. The iterative aggregation-disaggregation method

The hierarchical model comprises a lower level at which the two phase server resources are modelled in a detailed way, and a higher level at which the interaction between the resources is modelled as a separable queueing network model with queue length dependent service rates.

In the separable network the two phase server resource is approximated as a resource with a single service unit, a first-come first-served service discipline and queue length dependent service rates. The service demand is assumed to be exponentially distributed with unit mean. If resource n is a two phase server, the service rates $\mu_n(k)$, $k=1,\ldots,K$, are set to

$$\mu_n(k) = \frac{p(k,2)}{p(k,1)+p(k,2)} \frac{1}{w_n} . \tag{5.3.16}$$

The idea behind this relation is as follows. The lower level model yields for resource n a set of limiting probabilities $p(k,f)$, where $k=0,\ldots,K$ and $f=1,2$. Thus, the first part of the right hand side of (5.3.16) denotes the conditional probability that the resource is busy with phase two provided there are k customers present at resource n. The second part denotes the rate at which customers are leaving the resource in that situation.

The lower level model is analysed with a set of state dependent instream rates which follow from the analysis of the higher level model.
The rates of the two phases are $\nu=v_n^{-1}$ and $\mu=w_n^{-1}$ respectively. The instream rates $\lambda(k)$ are for $k=0,\ldots,K-1$ given by

$$\lambda(k) = T_n(K-k) , \tag{5.3.17}$$

where $T_n(K-k)$ is defined by (5.2.9).

The iteration between the two hierarchical levels yields approximations for the relevant performance characteristics at the distinct resources. Such an iteration may be started with an analysis of the higher level model. For this initial network the service rate function at a two phase server resource n may be set to

$$\mu_n(k) = \frac{1}{v_n + w_n} . \tag{5.3.18}$$

In practice the iteration method converges fast and yields quite acceptable results. In Section 5.6 we give some numerical examples.

5.4. Mean value analysis extensions

5.4.1. Introduction

In this section we discuss extensions of the MVA algorithm. The mean value relations are extended to account for the fact that the service disciplines at two phase server resources violate the separability conditions.
The adjustments are based on a mean value analysis of a single two phase server resource. This approach yields in a relatively simple way appealing intuitive approximation methods that are easy to implement in existing MVA algorithms.

In Subsection 5.4.2 the mean value analysis of the two phase server is dicussed. The implementation of the ideas in the MVA algorithm is covered in Subsection 5.4.3. Both a purely recursive and an iterative variant are presented.

5.4.2. Mean value analysis of a two phase server

Consider the following model. Customers arrive at a single service unit as a Poisson process with rate λ. The service discipline is first-come first-served. The service demand splits into two independent and exponentially distributed phases with means v and w respectively. The first phase is preparatory. The second phase can be executed only if the customer is present.

We derive relations between the expected residence time S of a customer in the queueing system and the expected number of customers in the system L. The derivation is based on the three important results: 1. the property that "Poisson Arrivals See Time Averages" (PASTA), 2. the expected residual life time formula and 3. Little's formula.
The property PASTA, Poisson Arrivals See Time Averages, has been used for long time in queueing system analysis. Only recently, a rigorous proof of the property has been given in Wolff [1982].
The expected residual life time formula is a standard result from renewal theory, cf. Kleinrock [1975:1] or Heyman and Sobel [1982]. For a queueing system it may be used in the following way: the expectation of the residual service demand of a customer equals

$(\sigma^2+w^2)/2w$. Here, w and σ denote the expectation and variance of the service demand distribution respectively.
Little's formula, cf. Subsection 3.3.3, relates the expected residence time S, the expected number of customers L and the throughput λ of a queueing system as $L=\lambda S$.

The first relation between S and L is provided by Little's formula

$$L = \lambda S . \tag{5.4.1}$$

The second relation matches the expected residence time of a newly arriving customer with the expected total amount of work to be executed with this customer being in the system.
Due to property PASTA, the arriving customer finds an expected number of L customers in front of it. The expected service demand of the customers equals $v+w$. So, roughly, the expected residence time equals the sum of the expected waiting time $L(v+w)$ and its own expected service demand $v+w$.
However, we have to amend for two obvious mistakes. In the first place, the newly arriving customer may find the service unit already busy with the second phase of the service demand of a customer. In the second place, the newly arriving customer may find the service unit idle. This implies that the customer is the first one in a busy period and that its preparatory phase has been completed in the preceding idle period.

Note that λv and λw denote the fractions of time that the service unit is busy with phase one and two respectively. With probability λv the newly arriving customer sees the service unit busy executing the first phase and so the expected residual service demand equals $v+w$. With probability λw the service unit is busy with the second phase and so the expected residual service demand equals w instead of $v+w$. With probability $1-\lambda v-\lambda w$ the service unit is idle when a customer arrives and the expected remaining service demand equals w instead of $v+w$.

When these observations are taken into account, we obtain the following relation for the expected residence time

$$S = (L + 1)(v+w) - \lambda wv - (1 - \lambda(v+w))v , \tag{5.4.2}$$

which may be rewritten as

$$S = L(v+w) + \lambda vv + w . \tag{5.4.3}$$

Relations (5.4.1) and (5.4.3) yield explicit expressions for the expected residence time and the expected number of customers in the two phase server system. In the next subsection we shall see how the reasoning may be used to approximate queueing network models with two phase servers.

5.4.3. Mean value analysis extensions

In Section 5.2 we have introduced the queueing network model with two phase servers. The basic queueing network model is separable. The MVA algorithm has been recapitulated.

The relations for the throughputs and the expected numbers of customers may be viewed as consequences of Little's formula and thus have a wide applicability. The expected residence time relations are consequences of an arrival theorem and cannot be used in this form for the analysis of a non-separable queueing network model.
However, we may use the idea of the arrival theorem in combination with the mean value anlysis of a two phase server as described in the preceding subsection to obtain approximate relations for the expected residence times.

The idea is to adopt the arrival theorem and to assume that in the non-separable queueing network model a customer sees upon a jump moment the system as if in equilibrium with itself removed. So, the relations of Section 5.2 remain intact at separable resources, though they are no longer exact as the network is no longer separable. At a two phase server resource, say n, the following adjustment of (5.4.3) is used

$$S_n(K) = L_n(K-1)(v_n + w_n) + \Lambda_n(K-1)v_n^2 + w_n \ . \tag{5.4.4}$$

The resulting adjusted MVA algorithm is again a recursive scheme and the implementation of (5.4.4) in an existing MVA algorithm is straightforward.

Note, that for $K=1$ the approximation for $S_n(1)$ equals w_n at a two phase server resource $n \in A$. Apparently, the approximation neglects the fact that the first phase has not necessarily been completed when the single customer that is in the system arrives.
A simple improvement to amend this obvious mistake is to use $\Lambda_n(K)$ rather than $\Lambda_n(K-1)$ in (5.4.4). This yields instead of (5.4.4)

$$S_n(K) = L_n(K-1)(v_n + w_n) + \Lambda_n(K)v_n^2 + w_n \ . \tag{5.4.5}$$

Observe, that now $\Lambda_n(K)$ has not been evaluated at the moment that $S_n(K)$ is to be evaluated.

A standard way to solve this problem is to start with an initial set of approximations for the throughputs $\Lambda_n(K)$. After an evaluation of the mean value analysis algorithm with this set of throughputs we have a new set of approximations at our disposal. This successive approximation method converges in all practical situations, but we have not investigated the method in detail.

5.5 An iterative approximation method

5.5.1. Introduction

In this section an iterative approximation method is introduced and analysed. The basic idea is to construct a separable queueing network model which approximates the original network model with two phase server resources. The parameters of the separable queueing network model are iteratively improved. The improvement is based on mean value arguments.
The method is introduced in Subsection 5.5.2. It is representative for a large class of iterative approximation methods. The iteration may be viewed as a successive approximation method for the determination of a fixed point of a continuous operator on a closed and convex subset of the n-dimensional Euclidean space. In Subsection 5.5.3 properties of the iteration method are studied for a model with only one two phase server. It is proved that for a broad class of queueing network models the method yields a unique fixed point.

5.5.2. The iterative approximation method

Consider a resource $n \in A$ being a special type of two phase server. The service demand at such a resource consists of two independent exponentially distributed phases with means v_n and w_n respectively. The effect of the deviating service discipline will be that some customers experience only a service demand w_n whereas others have the full demand $v_n + w_n$.
A natural approximation method seems to be to construct a separable queueing network model with at resource n an average service demand $\tilde{w}_n$ being a weighted average of v_n and w_n, namely

$$\tilde{w}_n = (1 - a_n)v_n + w_n \ , \tag{5.5.1}$$

where a_n denotes the probability that an arriving customer finds its preparatory phase finished.

Finding a_n requires a rigorous analysis of the original model and that we just wanted to avoid. However, one might make a first guess, for instance $a_n = 0$ or $a_n = 1$, and try to improve on this guess after an evaluation of the associated separable queueing network model.

We may write a_n as

$$a_n = b_n c_n \ , \tag{5.5.2}$$

where b_n is the probability that an arriving customer is the first on in a busy period and c_n the probability that its preparatory phase has been completed in the preceding idle period.
Estimates for b_n and c_n may be constructed as follows. The probability b_n equals the

probability that an arriving customer sees no customers in front of it and thus may be approximated by

$$b_n = p_n(0,K-1)\,, \tag{5.5.3}$$

where it is tacitly assumed that the arrival theorem still holds.
The expected duration of an idle period, say I_n, is the quotient of the fraction of time during which no customers are at resource n, $p_n(0,K)$, and the expected number of busy periods per unit time, $\Lambda_n(K)p_n(0,K-1)$, i.e.

$$I_n = \frac{p_n(0,K)}{\Lambda_n(K)p_n(0,K-1)}\,. \tag{5.5.4}$$

Assuming that an idle period is exponentially distributed yields for c_n

$$c_n = \frac{I_n}{I_n + v_n}\,. \tag{5.5.5}$$

So, suppose that we have a guess $a_n^{(i)}$. The new guess for a_n may now be introduced as

$$a_n^{(i+1)} = \frac{p_n^{(i)}(0,K-1)\,p_n^{(i)}(0,K)}{p_n^{(i)}(0,K) + \Lambda_n^{(i)}(K)p_n^{(i)}(0,K-1)v_n}\,, \tag{5.5.6}$$

where the values at the right hand side are obtained from a mean value analysis of a separable queueing network model with at a two phase server resource n an expected service demand $\tilde{w}_n^{(i)}$ given by

$$\tilde{w}_n^{(i)} = (1 - a_n^{(i)})v_n + w_n\,. \tag{5.5.7}$$

This completes the description of the iterative method. In Section 5.6 some numerical examples are given and the accuracy is compared with that of the methods described in the preceding sections. In the next subsection we analyse the iteration method for a system with exactly one two phase server.

5.5.3. Existence, uniqueness and convergence

This subsection is concerned with the analysis of the iterative approximation method which has been introduced in the preceding subsection. It is assumed that there is exactly one two phase server in the network.
The method will be formulated as a successive approximation method for the determination of a fixed point. The existence of a fixed point is a consequence of Brouwer's fixed point theorem. The convergence of the method and the uniqueness of the fixed point will be proved for a subclass of queueing network models.
For details on the presented analysis we refer to Van Doremalen and De Waal [1985].

Assume that resource n is the two phase server resource and that the visiting ratios have been normalized in such a way that $f_n = 1$. A non-linear operator $F:R \to R$ gives the relation between $\tilde{w}_n^{(i+1)}$ and $\tilde{w}_n^{(i)}$ as $\tilde{w}_n^{(i+1)} = F(\tilde{w}_n^{(i)})$, cf. Relations (5.5.1) through

(5.5.7),

$$F(x) = \left[1 - \frac{p_n(0,K-1)p_n(0,K)}{p_n(0,K) + \Lambda_n(K)p_n(0,K-1)v_n}\right]v_n + w_n \ , \tag{5.5.8}$$

where $p_n(0,K-1)$, $p_n(0,K)$ and $\Lambda_n(K)$ are functions of x.

Applying the results of Lemma 5.1 (5.5.8) reduces to

$$F(x) = \left[1 - \frac{1}{G(K-1)} \frac{G^{[n]}(K-1)\,G^{[n]}(K)}{G^{[n]}(K) + v_n G^{[n]}(K-1)}\right]v_n + w_n \ . \tag{5.5.9}$$

Note that $G^{[n]}(K-1)$ and $G^{[n]}(K)$ do not depend on x and can be treated as constants in the analysis of the operator F. Furthermore, we may write $G(K-1)$ as a polynomial of degree $K-1$ in the variable x, cf. Lemma 5.1,

$$G(K-1) = \sum_{k=0}^{K-1} G^{[n]}(K-1-k)x^k \ . \tag{5.5.10}$$

Now the analysis of the operator F is straightforward. It is assumed that at least two customers are in the system, i.e. $K > 1$. For $K = 1$ the results are trivially true as is easily verified.
We first show that F has at least one fixed point at the interval $[w_n, v_+ w_n]$. Afterwards it is shown that for a large class of queueing network models the fixed point is unique.

The existence of a fixed point is an immediate result of the next lemma

Lemma 5.2

Assume that $K > 1$. Then the following three statements hold:

(i) F is analytic on the interval $[w_n, v_n + w_n]$,

(ii) F is monotonically increasing on $[w_n, v_n + w_n]$, and

(iii) $F(x) \in (w_n, v_n + w_n)$ for all $x \in [w_n, v_n + w_n]$.

Proof :

To prove (i) it is sufficient to observe that F is a rational polynomial without any poles on the interval $[w_n, v_n + w_n]$.

To prove (ii) observe that for $x \in [w_n, v_n + w_n)$ $G(K-1)$ is a monotone increasing function of x as the coefficients $G^{[n]}(k)$ are strictly positive, cf. (5.2.15).

To prove (iii) observe that $F(x)$ may be written as

$$F(x) = (1 - D(x))v_n + w_n \ , \tag{5.5.11}$$

where

$$D(x) = \frac{G^{[n]}(K-1)}{G^{[n]}(K-1) + \sum_{k=1}^{K-1} G^{[n]}(K-1-k)x^k} \frac{G^{[n]}(K)}{G^{[n]}(K) + v_n G^{[n]}(K-1)} \tag{5.5.12}$$

As a consequence we have that, for all $x \in [w_n, v_n + w_n]$,

$$0 < D(x) < 1 , \tag{5.5.13}$$

which completes the proof. ∎

To prove convergence of the successive approximation method and uniqueness of the fixed point we first transform the operator F. The analysis is reduced to the analysis of the iteration function in a queueing network model with only two resources: the two phase server and a complementary resource replacing the complement of the network, cf. the discussion in Section 5.2.

The service rates of the complementary resource are $T_n(k)$, $k = 1,...K$, where

$$T_n(k) = \frac{G^{[n]}(k-1)}{G^{[n]}(k)} . \tag{5.5.14}$$

That this transformation yields the desired results from

$$\prod_{i=K-k}^{K-1} T_n(i) = \frac{G^{[n]}(k-1-k)}{G^{[n]}(K-1)} , \tag{5.5.15}$$

which in combination with (5.5.9) and (5.5.10) yields

$$F(x) = \left[1 - \frac{1}{1 + \sum_{k=1}^{K-1} x^k \prod_{i=K-k}^{K-1} T_n(i)} \frac{1}{1 + v_n T_n(K)} \right] v_n + w_n . \tag{5.5.16}$$

So, F is the iteration function of a queueing network model with a two phase server and a complementary resource with queue length dependent service rates.

The following lemma prepares for a theorem on the convergence of the successive approximation method defined by the operator F.

Lemma 5.3

Let the following condition be satisfied for a given K

$$T_n(1) \leqslant T_n(2) \leqslant \cdots \leqslant T_n(K) . \tag{5.5.17}$$

Then

$$0 < \frac{dF(x)}{dx} \leqslant \frac{v_n T_n(k-1)}{1 + v_n T_n(k)} < 1 , \tag{5.5.18}$$

for all $k = 2,...,K$ and $x \in [w_n, v_n + w_n]$.

Proof : Cf. Theorem 4.3.2.1 in Van Doremalen and De Waal [1985]. ∎

The next Theorem is an immediate consequence of Lemmas 5.2 and 5.3.

Theorem 5.2

Let the condition given by (5.5.17) be satisfied. Then the operator F defines a contraction on the interval $[w_n, v_n + w_n]$ and $F(x) = x$ has exactly one solution in $(w_n, v_n + w_n)$.

Proof :

Let $x, y \in [w_n, v_n + w_n]$ with $x < y$. Then it is well known that, for some $x < z < y$,

$$F(x) - F(y) = \frac{dF(z)}{dz}(x - y), \tag{5.5.19}$$

From Lemma 5.3 we have

$$| F(x) - F(y) | \leqslant \frac{v_n T_n(K-1)}{1 + w_n T_n(K)} | x - y | . \tag{5.5.20}$$

From Lemma 5.2 it follows that $F(x) \in (w_n, v_n + w_n)$ for all $x \in [w_n, v_n + w_n]$ and so F is a contraction on $[w_n, v_n + w_n]$. This implies that the successive approximation method converges to a unique solution of the equation $F(x) = x$, cf. Ortega and Rheinboldt [1970: 5.1.3].

∎

So, we have proved the convergence of the iterative approximation method for those queueing network models for which the complementary resource satisfies the condition given by (5.5.17). This condition states that the service rate of the complementary resource is a non-decreasing function of the number of customers present. This seems to be a quite natural condition and we show that it holds for a large class of queueing network models.

Consider the separable queueing network model of Section 5.2. In Van Doremalen and De Waal [1985] it is shown that the condition (5.5.17) is satisfied if the throughput at each resource is a non-decreasing function of the number of customers in the system. The next theorem provides a sufficient condition for this to be true. It gives an monotonicity result which in itself is of importance in the analysis of queueing network models.

Theorem 5.3

Let at each resource $n \in \{1,...,N\}$ the service capacity function satisfy the condition

$$\mu_n(1) \leqslant \mu_n(2) \leqslant \cdots \leqslant \mu_n(K). \tag{5.5.21}$$

Then for all $n = 1,...,N$

$$\Lambda_n(1) \leqslant \Lambda_n(2) \leqslant \cdots \leqslant \Lambda_n(K). \tag{5.5.22}$$

Proof :

This monotonicity result has been proved in Van Doremalen and De Waal [1985]. Independently a similar result has been formulated and proved in Suri [1985]. These proofs are based explicitly on the product form solution and proceed by induction. Related results have been reported in Robertazzi and Lazar [1984].
For an interesting alternative proof we refer to Van der Wal [1985]. In this paper an appealing intuitive proof has been given as well.

∎

5.6. Numerical examples

In this section we discuss four small numerical examples to show the behaviour of the different approximation methods. The exact evaluation has been obtained by the solution of the corresponding set of equilibrium equations, examples 1 and 2, and the method described in Subsection 5.3.2, examples 3 and 4. The exact results are denoted by EXACT.
We have tested the four approximation methods suggested in the preceding sections. The first method is the iterative aggregation-disaggregation method which in the tables has been called the IAD method. The next two methods are the recursive and iterative MVA extensions which are denoted as MVE-R and MVE-I respectively. The last method is the iterative method with mean value arguments which is denoted as MVA-I.

From the results it appears that especially the iterative aggregation- disaggregation method and the iterative method with mean value arguments perform very well. The MVA extensions perform slightly worse. It is our belief that for larger systems the approximations will tend to be better, but further research is needed to support this tentative conclusion.

Example 1 : A cyclic system

The first example concerns a cyclic queueing system with three resources. The first resource in the cycle is a two phase server with expected service demands $v_1 = 1$ and $w_1 = 1$ respectively. The second and third resource in the cycle are first-come first-served resources with expected service demands w_2 and w_3 respectively. There are four customers in the system.
In the Tables 5.1 and 5.2 we have listed the exact and approximate results for the throughput at the two-phase server and the expected cycle time respectively. This is done for various w_2 and w_3.

Example 2 : A branching system

The second example concerns a branching system with three resources. The first resource is a two phase server with expected service demands $v_1 = 1$ and $w_1 = 1$ respectively. The second and third resource in the branching system are first-come first-served resources withexpected service demands w_2 and w_3 respectively. After a visit to the two phase

server with probability 0.5 a visit to the second resource follows and with probability 0.5 to the third resource. After a visit to the second or third resource the two phase server is visited again. There are four customers in the system.
In the Tables 5.3 and 5.4 the results are shown for the throughput at the two phase server and the expected cycle time respectively. This is done for various w_2 and w_3.

Example 3 : A two phase server and an infinite server

The third example concerns a cyclic system comprising a two phase server and an infinite server. The expected service demands of the two phases are $v_1 = 1$ and w_1 respectively. The expected service demand at the infinite server is w_2. There are K customers in the system.
In the Tables 5.5 and 5.6 we have listed the throughputs and expected residence times at the two phase server for various values of w_2 and K.

Example 4 : A two phase server and a first-come first-served server

The fourth example considers a cyclic system comprising a two phase server and a first-come first-served server. The expected service demands of the two phase are $v_1 = 1$ and $w_1 = 1$ respectively. The expected service demand at the first-come first-served server is w_2. There are K customers in the system.
Tables 5.7 and 5.8 show the throughputs and expected residence times at the two phase server for various values of w_2 and K.

w_2	w_3	EXACT	IAD	MVE-R	MVE-I	MVA-I
.25	.25	.500	.500	.537	.535	.500
2	2	.357	.352	.353	.352	.351
8	8	.124	.124	.124	.124	.124
2	8	.124	.124	.124	.124	.124

Table 5.1 : Throughput at the two-phase server (example 1).

w_2	w_3	EXACT	IAD	MVE-R	MVE-I	MVA-I
.25	.25	8.00	8.00	7.45	7.48	8.00
2	2	11.38	11.36	11.33	11.36	11.38
8	8	32.19	32.19	32.19	32.19	32.19
2	8	32.19	32.19	32.19	32.19	32.19

Table 5.2 : Expected cycle times (example 1).

w_2	w_3	EXACT	IAD	MVE-R	MVE-I	MVA-I
.25	.25	.500	.500	.532	.533	.500
2	2	.477	.478	.501	.497	.473
8	8	.194	.194	.194	.194	.194
10	1	.198	.199	.198	.198	.199
20	1	.100	.100	.100	.100	.100

Table 5.3 : Throughput at the two phase server (example 2).

w_2	w_3	EXACT	IAD	MVE-R	MVE-I	MVA-I
.25	.25	8.00	8.00	7.52	7.50	8.00
2	2	8.39	8.37	7.98	8.05	8.96
8	8	20.57	20.57	20.58	20.58	20.57
10	1	20.22	20.10	20.20	20.20	20.10
20	1	40.03	40.00	40.00	40.00	40.00

Table 5.4 : Expected cycle times (example 2).

w_2	K	EXACT	IAD	MVE-R	MVE-I	MVA-I
10	10	.497	.497	.466	.520	.495
2	2	.444	.444	.500	.470	.441

Table 5.5 : Throughput at the two phase server (example 3).

w_2	K	EXACT	IAD	MVE-R	MVE-I	MVA-I
10	10	10.12	10.12	8.91	9.23	10.20
2	2	2.50	2.50	2.00	2.26	2.54

Table 5.6 : Expected residence time at the two phase server (example 3).

w_2	K	EXACT	IAD	MVE-R	MVE-I	MVA-I
2	10	.466	.466	.521	.466	.463
2	2	.371	.371	.375	.372	.366

Table 5.7 : Throughput at the two phase server (example 4).

w_2	K	EXACT	IAD	MVE-R	MVE-I	MVA-I
2	10	10.34	10.34	9.19	8.99	9.92
2	2	2.35	2.35	2.00	2.16	2.37

Table 5.8 : Expected residence time at the two phase server (example 4).

6. PRIORITY QUEUEING NETWORK MODELS

6.1. Introduction

This chapter is concerned with the approximate analysis of queueing network systems with a priority schedule at one or more of the resources. We concentrate on the preemptive resume priority schedule which is a frequently used tool in the modelling of priority schedules in computer systems.
The exact analysis of priority schedules is notoriously difficult. Only for relatively small and simple models exact results and efficient computational procedures have been derived, cf. for instance Jaiswal [1968], Avi-Itzhak and Heyman [1973], Marks [1973], Neuts [1981:pp 298-300], Morris [1981], Veran [1984] and Van Doremalen [1984:2].
For larger and more complex models simulation techniques and analytical approximation methods have to be considered. We concentrate on the latter way-out and present a new and promising approximation method.

The analysis of priority queueing network models is of paramount interest as priority disciplines are a natural and frequently used tool for improving the performance of a queueing network. The introduction of a priority schedule has consequences for production oriented characteristics, such as utilizations, and demand oriented characteristics, such as expected residence times and throughputs. A priority schedule may be implemented to find a good balance between these characteristics, for instance by trying to optimize the utilization under certain constraints on the expected residence times.

A first example is a CP-terminal system comprising a central processor (CP) and a set of active terminals. The users at the terminals generate jobs to be executed by the CP. The CP gives preference to the jobs of certain terminals. The service discipline operates under a preemptive resume priority schedule.: if a job with a higher priority enters the CP, it interrupts the execution of a job with a lower priority instantaneously. The execution of this lower priority job is resumed as soon as no jobs with a higher priority are present anymore. Observe that in this way the execution of a job may be interrupted one or more times by what we shall refer to as busy periods of higher priority jobs.
Priority schedules in computer systems are studied in for example Avi-Itzhak and Heyman [1973], Kleinrock [1975:2] and Kameda [1984].
For the CP-terminal system relatively efficient computational procedures for the exact evaluation of important performance characteristics have been developed in Veran [1984] and Van Doremalen [1984:2]. The impact of the preemptive resume priority schedule on the utilization of the CP has for instance been studied in Van der Wal [1982].

A second example is a closed central server model with central processor units and a set of shared background memory devices or I/O's, cf. also Subsection 4.6.3. The performance of the system may be improved by the introduction of a preemptive resume priority schedule at the central processor units. In practical situations the priority schedule will be more sophisticated, but the analysis of a comparitively simple schedule may provide

insight in the operation of the queueing network system under more complicated schedules, cf. Chandy and Sauer [1978].

Separable queueing network models are an important tool in the analysis of large and complex queueing network systems. Regrettably, many realistic problems either cannot be modelled as separable queueing network models or the models are very large and therefore intractable by the standard evaluation methods.
Queueing network systems with a priority schedule at some of the resources are examples of realistic and important problems which lead to non-separable and very large models.
In Sections 2.4 and 2.5 two approaches have been discussed for the approximate analysis of queueing network systems. These offer themselves as immediate candidates for the approximate analysis of queueing network models with priority schedules.

The first approach is based on the idea to approximate the non-separable queueing network model by a separable model. In this approach two lines may be discerned.
A first line uses ideas based on a parametric analysis, cf. Chandy, Herzog and Woo [1975:1] and [1975:2]. The usual idea is to perform an exact analysis of a relatively detailed but small priority queueing system, where the complement of the network has been aggregated in a single complementary resource. A further reduction in the complexity may be achieved by introducing an aggregation of priority levels as well. Typical examples in this line are the methods described in Sauer and Chandy [1975], Chow and Yu [1983] and Neuse and Chandy [1983].
A second line started in Reiser [1976] and Sevcik [1978] with the development of the virtual server or shadow-CPU approximation. The main idea is to decompose a single priority queue in a set of parallel queues, each appointed for the execution of jobs of a particular priority level. The service rates or service demands at these resources are adjusted to account for the influence of the customers of the other priority levels. Improvements have been suggested in Schmitt [1984] and Kaufmann [1984].

The second approach is based on adjustments and extensions of the MVA algorithm. For the approximate analysis of priority schedules we refer to Bard [1979], Bryant, Krzesinski and Teunissen [1983] and Chandy and Lakshmi [1983]. Typical for these methods is that they are based on the MVA approach as sketched in Chapter 3 and the analysis of an M/M/1//PR-queue, i.e. a single server system with Poisson arrival processes, exponential service demand distributions and a preemptive resume priority schedule.

In this chapter we follow the second approach and present a new and promising approximation method: the service completion time approximation.

The basic queueing network model is a separable queueing network model with constant service rates, cf. Subsection 3.3.5. For reasons of presentation the model and its MVA algorithm are briefly recapitulated in Section 6.2. At some of the resources a preemptive resume priority schedule is introduced. In Section 6.4 we present the new approximation method which is based on the MVA approach and the mean value analysis of M/G/1-priority queues. This latter analysis is reviewed in Section 6.3.

As our method is based on the MVA analysis of closed multichain queueing network models, where the chains and priority levels have to be identified, the computational complexity and storage requirements of the method are similar to that of the original MVA algorithm. This prohibits the exact evaluation for larger numbers of priority levels. In Section 6.5 the use of the approximation methods that have been presented in Chapter 4 is considered. The concentration is on the use of the Schweitzer method and the global aggregation method, both with a first order depth improvement.

In Section 6.6 we present an exact procedure for the evaluation of important performance measures in the CP-terminal system. The material is based on Van Doremalen [1984:2] and [1984:3].

In Section 6.7 some numerical examples are presented to illustrate the effectiveness of the approximation methods. For a detailed analysis of numerical experiences we refer to Van Doremalen, Wessels and Wijbrands [1985].

6.2. The basic closed multichain queueing network model

The basic model is a separable and closed multichain queueing network model. It is a special case of the model discussed in Subsection 3.3.5. For reasons of presentation we shortly recapitulate the model and its mean value analysis.

The network comprises N resources. We allow for three service disciplines: first-come first-served (FCFS), processor sharing (PS) and infinite server (IS). It is assumed that the service rates at each of these resources are fixed and normalized to unity.
There are R closed customer chains. The K_r customers of chain r follow a Markov routing with visiting ratios $f_{n,r}$ at a resource n. The service demands of the customers of chain r at resource n are stochastically independent random variables with mean $w_{n,r}$. The service demands at a FCFS resource are exponentially distributed with a common mean for all chains.

The studied steady state system characteristics are

$S_{n,r}(K)$ expected residence time of customers of chain r at resource n,

$\Lambda_{n,r}(K)$ throughput of customers of chain r at resource n and

$L_{n,r}(K)$ expected number of customers of chain r at resource n,

where the argument K expresses the dependency on the population vector.
The MVA approach, as sketched in Subsection 3.3.5, leads to a recursive procedure for the evaluation of the characteristics.

For all population vectors $\boldsymbol{k}$ in the range of (0,...,0) through $(K_1, \ldots, K_R)$ a set of mean value relations has to be evaluated.
The expected residence time of a chain r customer at a FCFS or PS resource is given by

$$S_{n,r}(\boldsymbol{k}) = \sum_{i=1}^{R} L_{n,i}(\boldsymbol{k}-\boldsymbol{e}_r)w_{n,r} + w_{n,r} \tag{6.2.1}$$

and at an IS resource by

$$S_{n,r}(\boldsymbol{k}) = w_{n,r} \ . \tag{6.2.2}$$

The throughput of customers of chain r at resource n is given by

$$\Lambda_{n,r}(\boldsymbol{k}) = \frac{f_{n,r}k_r}{\sum_{m=1}^{N} f_{m,r}S_{m,r}(\boldsymbol{k})} \ . \tag{6.2.3}$$

The expected number of chain r customers at resource n is given by

$$L_{n,r}(\boldsymbol{k}) = \Lambda_{n,r}(\boldsymbol{k})S_{n,r}(\boldsymbol{k}) \ . \tag{6.2.4}$$

6.3. A mean value analysis of M/G/1//PR-queues

Consider a queueing system with a single service unit and a fixed service rate which is normalized to unity. Customers arrive as R independent Poisson streams with rates λ_r, $r=1,\ldots,R$. The service demands of the customers of stream r are stochastically independent random variables with distribution functions G_r with mean w_r and variance σ_r^2. The customers of a given stream are served in order of arrival, but between the streams a preemptive resume priority schedule governs the service discipline. This schedule is such that customers of stream i have a higher priority level than customers of stream r if $i<r$. If a higher priority customer arrives during the service of a lower priority customer, the service of the latter customer is instantaneously interrupted in favor of the new customer. The serving of the lower priority customer is resumed as soon as no higher priority customers are present anymore.
In such a system, which shall be referred to as an M/G/1//PR queue, the steady state performance measures may be derived by means of a mean value reasoning. This reasoning is based on the following three results for the stochastic process under consideration: 1. Poisson arrivals see time averages, 2. the expected residual life time formula, and 3. Little's formula, cf. Section 5.3 for more details.
For a more detailed analysis we refer to Gaver [1962], Cobham [1954], Takacs [1964], Wolff [1970] and Stidham [1972].

The system is characterized by

R	the number of priority levels,
λ_r	the instream rate of customers with priority level r,
w_r	mean service demand of a customer of steam r, and
σ_r^2	variance of the service demand distribution of a customer of stream r.

The following steady state characteristics play a role

S_r expected residence time of a stream r customer in the system,

C_r expected service completion time of a stream r customer, i.e. the expected length of the time interval which passes between the moment that the customer is taken into service for the first time and the moment that its service has been completed,

L_r expected number of stream r customers in the system, and

Q_r probability that a stream r customer is in the service completion phase.

It is assumed that $\lambda_1 w_1 + \ldots + \lambda_R w_R < 1$. This implies that the queueing process is not degenerated, since the total amount of work offered to the system does not exceed the capacity of the service unit.

We discuss a derivation of direct relations between these system characteristics based on a mean value analysis reasoning. The technique is very similar to the one described in Section 5.3.

The reasoning starts with a relation for the expected residence time of customers of stream r, $r = 1,\ldots,R$. Since Poisson arrivals see time averages, a stream r customer sees upon its arrival at the priority queue on the average L_i customers of stream i, $i = 1,\ldots,R$, present and with probability Q_i it finds a stream r customer in the service completion phase. The expected remaining service demand of the $L_i - Q_i$ customers which are not in the service completion phase, equals the mean of the distribution function G_i, i.e. w_i. The expected remaining service demand, say m_i, of a customer of stream i being in the service completion phase equals the expectation of the limiting distribution of the remaining service demand as seen by a random observer (Poisson arrivals see time averages), i.e.

$$m_i = \frac{\sigma_i^2 + w_i^2}{2w_i} . \tag{6.3.1}$$

A customer of stream r has to wait for the service completion of the customers of the streams $i = 1,\ldots,r$ which are in the system upon its arrival. Furthermore, during its residence in the system new customers of the higher priority levels $i = 1,\ldots,r-1$ are arriving at a rate λ_i. These customers are to be served in front of the customer of stream r as well. The expected residence time thus atains the following form

$$S_r = \sum_{i=1}^{r} (L_i - Q_i) w_i + \sum_{i=1}^{r} Q_i m_i + w_r + \sum_{i=1}^{r-1} \lambda_i S_r w_i . \tag{6.3.2}$$

To obtain an expression for the expected number of customers of stream r we apply Little's formula

$$L_r = \lambda_r S_r . \tag{6.3.3}$$

To evaluate the (limiting) probability that a customer of stream r is in the service completion phase we observe that the expected length of the service completion phase satisfies the following equation

$$C_r = w_r + \sum_{i=1}^{r-1} \lambda_i C_r w_i \ , \tag{6.3.4}$$

which yields

$$C_r = \frac{w_r}{1 - \sum_{i=1}^{r-1} \lambda_i w_i} \ . \tag{6.3.5}$$

Applying Little's formula yields

$$Q_r = \lambda_r C_r \ . \tag{6.3.6}$$

Introduce, for $i, r = 1,\ldots,R$, $W_{i,r}$ and $M_{i,r}$ as

$$W_{i,r} = \frac{w_i}{1 - \sum_{j=1}^{r-1} \lambda_j w_j} \tag{6.3.7}$$

and

$$M_{i,r} = \frac{m_i}{1 - \sum_{j=1}^{r-1} \lambda_j w_j} \ . \tag{6.3.8}$$

Then, we may rewrite (6.3.2) as

$$S_r = \sum_{i=1}^{r} (L_i - Q_i) W_{i,r} + \sum_{i=1}^{r} Q_i M_{i,r} + W_{r,r} \ . \tag{6.3.9}$$

The above sketched relations constitute a recursive scheme for the evaluation of the performance characteristics in a M/G/1//PR-queue. One may even give a closed form expression for the expected residence times. By induction the following lemma may be shown. For similar results see for instance Takacs [1964] and Wolff [1970].

Lemma 6.1

For $r = 1,\ldots,R$

$$S_r = \frac{\sum_{i=1}^{r} \lambda_i w_i m_i}{(1 - \sum_{j=1}^{r-1} \lambda_j w_j)(1 - \sum_{j=1}^{r} \lambda_j w_j)} + \frac{w_r}{1 - \sum_{j=1}^{r-1} \lambda_j w_j} \ . \tag{6.3.10}$$

∎

The Relations (6.3.7) through (6.3.9) form the basis for the approximation methods to be discussed in the next section.

6.4. A service completion time approximation

Let us consider the closed queueing network model of Section 6.2. At some of the resources a preemptive resume priority schedule is implemented. For reasons of presentation it is assumed that the priority levels correspond with the customer chain numbers. This implies that at all the resources with a priority schedule R priority levels are discerned which correspond with the R closed customer chains. Furthermore, at all priority resources the customers of chain 1 have the highest priority and of chain R the lowest. It is straightforward to extend the reasoning of this section to more general schedules.

As we have seen in Section 6.2 the model without priority queues may be evaluated by means of the recursive MVA algorithm. For the model with priority queues the relations for the throughput (6.2.3) and the expected numbers of customers (6.2.4) remain valid as these are based on applications of Little's formula. The relation for the expected residence time (6.2.1) will be violated as the arrival theorem will no longer hold.

A natural way to design an approximation method offers itself. One has to adjust the relation for the expected residence times in such a way that the violation of the arrival theorem and the introduction of the alternative service discipline are compensated. These adjustments may be founded on a study of the behaviour of a priority queue in a more simple environment and on the formulation of an approximating arrival theorem.
The simple environment has been sketched in the preceding section. As an approximating arrival theorem we simply extend the MVA approach to the non-separable queueing network model: the limiting distribution at arrival epochs of customers of chain (priority level) r equals the limiting distribution of the same system with one customer of chain r removed.
Note that the latter assumption implies that the expected residence time relations for FCFS and PS resources are given by the Relation (6.2.1). For the formulation of an approximating expected residence time relation we, therefore, may concentrate on resources with a priority schedule.

Consider a resource n with a preemptive resume priority schedule. Combining Relation (6.3.9) for the expected residence time of a customer of priority level r at an M/G/1//PR queue and the approximating arrival theorem, we formulate the following approximating expression for the expected residence time of a customer of chain r at a preemptive resume resource n

$$S_{n,r}(\boldsymbol{K}) = \sum_{i=1}^{r} (L_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r) - Q_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r))W_{n,i,r}(\boldsymbol{K}) + \sum_{i=1}^{r} Q_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r)M_{n,i,r}(\boldsymbol{K}) + W_{n,r,r}(\boldsymbol{K}) . \tag{6.4.1}$$

Combining (6.3.5) and (6.3.6) yields the following approximation for the probabilities $Q_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r)$

$$Q_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r) = \frac{\Lambda_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r)w_{n,i}}{1 - \sum_{j=1}^{i-1} \Lambda_{n,j}(\boldsymbol{K}-\boldsymbol{e}_r)w_{n,j}} . \tag{6.4.2}$$

This relation differs from (6.3.6) just by the addition of the resource index n and the dependency on the population vector K.

First guesses for $W_{n,i,r}(K)$ and $M_{n,i,r}(K)$ follow from (6.3.7) and (6.3.8), namely

$$W_{n,i,r}(K) = \frac{w_{n,i}}{1 - \sum_{j=1}^{r-1} \Lambda_{n,j}(K - e_r) w_{n,j}} \tag{6.4.3}$$

and

$$M_{n,i,r}(K) = \frac{m_{n,i}}{1 - \sum_{j=1}^{r-1} \Lambda_{n,j}(K - e_r) w_{n,j}} . \tag{6.4.4}$$

The idea to base the approximation on the values for the population vector $K - e_r$ rather than for K must be seen in the light of the arrival theorem. If we assume the arrival theorem to hold (approximately) in the network model with priority queues, it might be argued that the number of interrupts during the residence of a given customer is determined by the situation where this customer is not in the system.
As an aside it should be noted that in practical situations the removal of a low priority customer will hardly influence the throughput of higher priority customers. Then, the argument that the use of the population vector $K - e_r$ leads to a purely recursive scheme, whereas the use of population vector K does not necessarily implies this property, may be a decisive argument to use the dependency on $K - e_r$.
For exponentially distributed service demands the above sketched scheme has been mentioned e.g. in Bryant, Krzesinski and Teunissen [1983]. An improvement has been suggested in Chandy and Lakshmi [1983]. We shall refer to the approximation as the MVA approximation.

The idea behind Relation (6.4.2) is that there is a constant instream of customers of priority level i during the sojourn of a customer of priority level r, where $i = 1,...,r-1$. The method is based on an approximation of this instream rate. For open queueing network systems as the M/G/1//PR queue this reasoning is sound enough.
However, the fact that in a closed queueing network model the populations are finite introduces a complication. Especially the instream rate of customers of the levels $i+1,...,r-1$ during the service completion of customers from the levels $1,...,i$ seems to be grossly overestimated by (6.4.3) and (6.4.4). We therefore suggest to approximate $W_{n,i,r}(K)$ and $M_{n,i,r}(K)$ by

$$W_{n,i,r}(K) = \frac{w_{n,i}}{1 - \sum_{j=1}^{i-1} \Lambda_{n,j}(K - e_r w_{n,j}} \tag{6.4.5}$$

and

$$M_{n,i,r}(K) = \frac{m_{n,i}}{1 - \sum_{j=1}^{i-1} \Lambda_{n,j}(K - e_r) w_{n,j}} . \tag{6.4.6}$$

Observe that these expressions are approximations for the expected service completion time of a customer of chain i and the expected remaining service completion time, cf. Relation (6.3.5). The approximation method suggested by the Relations (6.4.5) and (6.4.6) shall therefore be referred to as the SCT or service completion time approximation.

6.5. Closed multichain queueing network approximations

6.5.1. Introduction

The approximation methods suggested in the preceding section use the recursive MVA algorithm for closed multichain queueing network models. As we have seen, the computational complexity and storage requirements of this algorithm prohibit an exact evaluation for larger values of $R, K_1, \ldots, K_R$. A number of approximation methods has been proposed in Chapter 4. The Schweitzer method and the global aggregation method appeared to be efficient and accurate tools for the evaluation of approximations for some important performance characteristics. Especially in combination with a first order depth improvement the methods performed quite satisfactory. In this section we discuss the use of these methods for the approximate analysis of queueing network models with preemptive resume priority queues.
In order not to obscure the line of argument, the service demands at the priority queues are assumed to be exponentially distributed. As a consequence the expected residence time relations attain a more simple form. The extension to general service demand distributions is straightforward.

For the sake of completeness let us formulate the expected residence time relation at a preemptive resume priority queue with exponential service demand distributions. It is easily verified that (6.4.1) reduces to

$$S_{n,r}(\boldsymbol{K}) = \sum_{i=1}^{r} L_{n,i}(\boldsymbol{K}-e_r) W_{n,i,r}(\boldsymbol{K}) + W_{n,r,r}(\boldsymbol{K}) . \tag{6.5.1}$$

We study the MVA approximation and the SCT approximation as suggested by (6.4.3) and (6.4.5) respectively.

6.5.2. The Schweitzer method

The Schweitzer method is founded on the idea of removing the recursion from the MVA algorithm and to concentrate on the evaluation of the performance characteristics at the population vector $\boldsymbol{K}$, cf. Section 4.3.
To construct a non-recursive set of mean value relations at the population vector $\boldsymbol{K}$ the following approximations $L_{n,i,r}$ are introduced for $L_{n,i}(\boldsymbol{K}-e_r)$

$$L_{n,i,r} = \begin{cases} L_{n,i}(\boldsymbol{K}) & , i \neq r \\ \dfrac{K_r - 1}{K_r} L_{n,r}(\boldsymbol{K}) & , i = r \end{cases} \tag{6.5.2}$$

Furthermore, it is suggested to approximate $\Lambda_{n,i}(\boldsymbol{K}-\boldsymbol{e}_r)$ by $\Lambda_{n,i,r}$, where,

$$\Lambda_{n,i,r} = \Lambda_{n,i}(\boldsymbol{K}) . \tag{6.5.3}$$

The mean value relations at the population vector $\boldsymbol{K}$ then transform in a set of non-linear equations for the performance characteristics at the population vector $\boldsymbol{K}$. Omitting the argument $\boldsymbol{K}$ we may write these relations as follows.

At a first-come first-served or processor sharing resource n the expected residence time relation for a customer of chain r is

$$S_{n,r} = \sum_{i=1}^{R} L_{n,i,r} w_{n,r} + w_{n,r} \tag{6.5.4}$$

and at an infinite server resource n

$$S_{n,r} = w_{n,r} . \tag{6.5.5}$$

As an approximation for the expected residence time at a resource with a preemptive resume priority schedule we obtain instead of (6.5.1)

$$S_{n,r} = \sum_{i=1}^{r} L_{n,i,r} W_{n,i,r} + W_{n,r,r} \tag{6.5.6}$$

For the values $W_{n,i,r}$ the approximations proposed in the preceding section are used in combination with (6.5.3). The first idea, the MVA approximation, leads to

$$W_{n,i,r} = \frac{w_{n,i}}{1 - \sum_{j=1}^{r-1} \Lambda_{n,j,r} w_{n,j}} , \tag{6.5.7}$$

wheras the second idea, the SCT approximation, leads to

$$W_{n,i,r} = \frac{w_{n,i}}{1 - \sum_{j=1}^{i-1} \Lambda_{n,j,r} w_{n,j}} . \tag{6.5.8}$$

The relation for the throughput of chain r customers at resource n is

$$\Lambda_{n,r} = \frac{f_{n,r} K_r}{\sum_{m=1}^{N} f_{m,r} S_{m,r}} . \tag{6.5.9}$$

The relation for the expected number of chain r customers at resource n is

$$L_{n,r} = \Lambda_{n,r} S_{n,r} . \tag{6.5.10}$$

This set of non-linear equations for the approximate characteristics may be solved in a standard way by successive approximations. Numerical experiments have shown the method to converge in all situations considered. We have implemented the Schweitzer approximation in combination with a first order depth improvement and have found the

method to be a very reliable and accurate tool for the approximation of the relevant performance characteristics. Especially the method based on the service completion time approximation yields very good results.

6.5.3. The global aggregation method

The global aggregation method, cf. Section 4.5, may be extended in a straightforward way to include approximations for preemptive resume priority resources.
This method consists of a recursive scheme running through all integer values k in the range $0,\ldots,T$, where

$$T = \sum_{r=1}^{R} K_r \ . \tag{6.5.11}$$

The expected residence time for a chain r customer at a first-come first-served or processor sharing resource n is given by

$$S_{n,r}(k) = \sum_{i=1}^{R} L_{n,i}(k-1)w_{n,r} + w_{n,r} \tag{6.5.12}$$

and at an infinite server resource n by

$$S_{n,r}(k) = w_{n,r} \ . \tag{6.5.13}$$

The expected residence time at a preemptive resume priority resource n may be approximated by the following variant of (6.5.1)

$$S_{n,r}(k) = \sum_{i=1}^{r} L_{n,i}(k-1)W_{n,i,r}(k-1) + W_{n,r,r}(k-1) \ . \tag{6.5.14}$$

For the MVA approximation the value $W_{n,i,r}(k-1)$ is put to

$$W_{n,i,r}(k-1) = \frac{w_{n,i}}{1 - \sum_{j=1}^{r-1} \Lambda_{n,j}(k-1)w_{n,j}} \tag{6.5.15}$$

and for the SCT approximation to

$$W_{n,i,r}(k-1) = \frac{w_{n,i}}{1 - \sum_{j=1}^{i-1} \Lambda_{n,j}(k-1)w_{n,j}} \ . \tag{6.5.16}$$

The throughput of customers of chain r at resource n is approximated by

$$\Lambda_{n,r}(k) = \frac{K_r}{T} \frac{f_{n,r}k}{\sum_{m=1}^{N} f_{m,r}S_{m,r}(k)} \tag{6.5.17}$$

and the expected number of chain r customers at resource n by

$$L_{n,r}(k) = \Lambda_{n,r}(k) S_{n,r}(k) . \tag{6.5.18}$$

The above described scheme is purely recursive and very efficient: the number of recursion steps is linear in the total number of customers instead of exponential in the number of customer chains. In the discussion of the numerical results we have considered a first-order depth improvement of the global aggregation method. It should be observed that the last step of such a depth improvement may be performed by means of the relations suggested in the Sections 6.2 and 6.4.
The numerical results are very promising and justify one of the main conclusions of our research: for the approximate analysis of closed queueing network models the use of purely recursive methods yields appealing, efficient and accurate approximation methods.

6.6. The CP-terminal system with preemptive resume priorities

6.6.1. Introduction

This section deals with the exact analysis of a priority queueing system which models a real life system comprising a set of terminals coupled with a single central processor system, a so-called CP-terminal system.

There are R groups of terminals, numbered $r = 1,...,R$. The K_r terminals of group r have independent and exponentially distributed thinktimes with parameter λ_r. The processing times of their jobs at the CP are independent and have a distribution function G_r with mean w_r. At the CP a preemptive resume priority schedule is implemented. The jobs of a terminal from group i have a higher priority than the jobs from group r if $i < r$.
For jobs from terminals of the same terminal group one may assume any work-conserving service discipline, for example first-come first-served, processor sharing or preemptive resume priorities, cf. Wolff [1970].
The user at the terminal thinks only if he has no job at the CP. Otherwise he is waiting for a response from the CP. So each user has at most one job at the CP at a time.

Our main interest is in the consequences of the priority schedule on the utilization of the CP and the expected response times of the jobs of the various priority levels. Thus it is our main purpose to evaluate global performance measures.

The system is a queueing system with a single service unit, finite Poisson sources, a preemptive resume priority service discipline and general processing time distributions. For the system with infinite Poisson sources we have described a mean value analysis in Section 6.3. For the system with finite Poisson sources, until recently, only complicated results in terms of Laplace-Stieltjes transforms were known. In Jaiswal [1968] an extensive study of such systems has been presented.
For the evaluation of utilizations and expected response times in a CP-terminal model with exponential think and processing times an exact algorithm has recently be presented in Veran [1984]. The analysis is based on a detailed study of the equilibrium equations of

the underlying continuous time Markov process.
We present a recursive algorithm for the evaluation of the utilizations and expected response times based on mean value arguments and properties of regenerative processes that can be recognised in the stochasic process describing the behaviour of the queueing system. It is shown that, for exponentially distributed processing times, the resulting scheme coincides with the one presented in Veran [1984]. Our analysis extends to the situation where the processing times have a general distribution in case a terminal group consists of a single terminal. For group with more than one terminal the exponentiality of the processing time distribution is essential.

This section is organized as follows. In Subsection 6.6.2 the recursive scheme for the CP-terminal system with general processing time distributions and a single terminal per group is derived.
In Subsection 6.6.3 it is shown how this analysis may be extended to cover systems with more than one terminal per group.
The recursive scheme induces a recursive algorithm for the evaluation of the main performance characteristics. In Van Doremalen [1984:3] an efficient algorithm has been presented to execute this recursive algorithm. In Subsection 6.6.4 we make some final remarks with respect to the analysis presented.

6.6.2. The CP terminal model with one terminal per group

In this subsection we discuss the exact analysis of a CP-terminal model with one terminal per group. Observe that this implies that the system comprises R terminals and R priority levels at the central processor. A recursive scheme is derived for the evaluation of the fractions of time the central processor is busy with jobs of a given terminal. The expected response times are simple functions of these quantities.

To get an intuition for the line of reasoning the following observations are made.
First, note that the first r terminals behave as if the last $R-r$ terminals did not exist. Obviously, the reverse is not true.
Secondly, note that the stochastic behaviour of terminal r may be analysed by considering a closed cyclic queueing system comprising two resources and a single customer which alternately visits these two resources. The service demand at one resource describes the thinktime at the terminal. The other resource models the execution of a job of the terminal. It behaves as a resource with a single service unit subject to breakdowns. The breakdowns are busy periods of higher priority jobs from the terminals $1,\ldots,r-1$.
These observations indicate the possibility of a recursive analysis based on an analysis of busy periods.

The basic recursive scheme

The evaluation of the performance measures is based on a mean value analysis of a special type of busy cycles. Before presenting the basic recursion the following terminology is

introduced to describe these busy cycles

busy r—period a busy period of jobs from the first r terminals at the central processor, i.e. an uninterrupted period of time during which the central processor is executing jobs from the first r terminals,

idle r—period an idle period with respect to jobs from the first r terminals at the central processor, and

busy r—cycle a busy cycle of the first r terminals, i.e. a combination of an idle and a busy r-period.

The basic observations are: 1. during a busy r-cycle at most one job from terminal r is executed and 2. a busy r-cycle forms a renewal cycle in the stochastic process describing the behaviour of the first r terminals. It should be noted that the start of a busy r-period is a regeneration point in this process as well.
Let us introduce the following notations:

Λ_r $\sum_{i=0}^{r} \lambda_i$, where $\lambda_0 = 0$ by definition,

u_r fraction of time the CP is executing jobs from terminal r,

U_r $\sum_{i=0}^{r} u_i$, where $u_0 = 0$ by definition, and

S_r expected response time of a job from terminal r at the CP, including service and waiting time.

Consider the situation that we have analysed the behaviour of the first $r-1$ terminals and that we are interested in the evaluation of u_r. One of the following three events occurs in a busy r-cycle: 1. a busy r-period starts with a job from terminal r, or 2. one job from terminal r enters during the busy r-period, or 3. during the busy r-period no job from terminal r is executed at all.
Studying the structure of a busy r-cycle and conditioning on the first two events, one may verify that

$$\frac{\lambda_r}{\Lambda_r} + \left[1 - \frac{\lambda_r}{\Lambda_r}\right] \pi_r \,, \tag{6.6.1}$$

is the probability that during a busy r-cycle a job from terminal r is executed, where π_r is to be interpreted as

π_r probability that during an arbitrary busy r-period a job from terminal r is executed, given that this period does not start with a job from terminal r.

The evaluation of the probabilities π_r, $r = 1,\ldots,R$, is studied later on. We now proceed with a mean value analysis of a busy r-cycle.

The expected number of busy r-periods per unit time equals the expected number of idle r-periods per unit time. The latter expectation is the quotient of the long run fraction of time that none of the first r terminals has a job at the CP, $1-U_r$, and the expected length of an idle r-period, Λ_r^{-1}. The long run fraction of time the CP is executing jobs from terminal r is the product of the expected number of busy r-periods per unit time, the probability that during a given busy r-period a job from terminal r is executed, and the expected processing time of a job from terminal r.
This yields, for $r=1,\ldots,R$,

$$u_r = \frac{1-U_r}{\Lambda_r^{-1}} \left[\frac{\lambda_r}{\Lambda_r} + (1 - \frac{\lambda_r}{\Lambda_r})\pi_r \right] w_r \ . \tag{6.6.2}$$

Inserting $U_r = U_{r-1} + u_r$ and $\Lambda_r = \Lambda_{r-1} + \lambda_r$ and solving for U_r yields for $r=1,\ldots,R$

$$U_r = \frac{U_{r-1} + (\lambda_r + \Lambda_{r-1}\pi_r)w_r}{1 + (\lambda_r + \Lambda_{r-1}\pi_r)w_r} \ . \tag{6.6.3}$$

With starting values $U_0=0$ and $\Lambda_0=0$ Relation (6.6.3) induces a recursive scheme for the evaluation of the fractions of time u_r that the CP is executing jobs from a specific terminal r if the probabilities π_r can be evaluated.

One may verify that the expected response time S_r of a job from terminal r at the CP is given by relation

$$S_r = \frac{w_r}{u_r} - \frac{1}{\lambda_r} \ . \tag{6.6.4}$$

Let us now concentrate on the evaluation of the probabilities π_r.

Evaluation of the probabilities π_r

The probability π_r depends on the think rate λ_r of the user at terminal r and on the think rates and processing time distributions of the users at the terminals $i=1,\ldots,r-1$. It does not depend on the processing time distribution of a job from the user at terminal r. The terminals $i=r+1,\ldots,R$ have no influence at all.
This leads to the introduction of auxiliary probabilities $\pi_r(\lambda)$ which, for r=1,...,R and $\lambda \geqslant 0$, are defined as

$\pi_r(\lambda)$ conditional probability that during a busy r-period at least one customer from a Poisson process with parameter λ arrives, if the period does not start with a job from terminal r.

For $r=1$ and $\lambda \geqslant 0$ we may initialize $\pi_1(\lambda)=0$. The probabilities π_r correspond with the probabilities $\pi_r(\lambda_r)$ for all $r=1,\ldots,R$.
The remainder of this subsection is devoted to the derivation of a recursive scheme for computing the probabilities $\pi_{r+1}(\lambda)$ for a given r in the range of $r=1,\ldots,R-1$. The method is based on the use of the structure of busy r-periods.

The first job in a busy r-period is from one of the terminals $i = 1,...,r$. Conditioning on the first job being either from the user at terminal r or from one of the users at the terminals $i = 1,...,r-1$, we obtain for all $\lambda \geqslant 0$

$$\pi_{r+1}(\lambda) = \frac{\lambda_r}{\Lambda_r} \Psi_r(\lambda) + \left[1 - \frac{\lambda_r}{\Lambda_r}\right] \Phi_r(\lambda), \tag{6.6.5}$$

where $\Psi_r(\lambda)$ and $\Phi_r(\lambda)$ should be interpreted as

$\Psi_r(\lambda)$ conditional probability that during a busy r-period at least one customer of a Poisson process with rate λ arrives, if it starts with a job from terminal r and

$\Phi_r(\lambda)$ conditional probability that during a busy r-period at least one customer from a Poisson process with parameter λ arrives, if it starts with a job from one of the terminals $1,...,r-1$, i.e. if it starts with a busy (r-1)-period.

Let us first have a closer look at the probabilities $\Phi_r(\lambda)$. For $r = 1$ and $\lambda \geqslant 0$ we may initialize with

$$\Phi_1(\lambda) = 0. \tag{6.6.6}$$

For $r = 2,...,R$ a busy r-period starts with a busy (r-1)-period if it does not start with a job from terminal r. At the end of a busy (r-1)-period one of the following three events has occured

1. with probability $\pi_r(\lambda)$ at least one customer from a Poisson process with parameter λ has arrived,

2. with probability $\pi_r(\lambda + \lambda_r) - \pi_r(\lambda)$ a job from terminal r has arrived and no customer from a Poisson process with parameter λ, and

3. with probability $1 - \pi_r(\lambda + \lambda_r)$ the end of the busy (r-1)-period coincides with the end of the busy r-period and no customer from a Poisson process with parameter λ has arrived.

Conditioning on the first two events we obtain

$$\Phi_r(\lambda) = \pi_r(\lambda) + [\pi_r(\lambda + \lambda_r) - \pi_r(\lambda)]\Psi_r(\lambda). \tag{6.6.7}$$

Next we concentrate on the evaluation of the probabilities $\Psi_r(\lambda)$. Here, for the first time the explicit processing time distribution play a role in the analysis. We first discuss the situation that the processing time of a job from terminal r is exponentially distributed. Afterwards the more complicated and somewhat different analysis for a general processing time distribution is given.

Exponentially distributed processing times

Suppose that the processing time of a job from terminal r is exponentially distributed with parameter μ_r. If the busy r-period starts with a job from terminal r, three events may occur: 1. the job is interrupted by an arriving higher priority customer, i.e. a busy (r-1)-period, 2. the job is interrupted by an arrival from a Poisson process with parameter λ, or 3. the job is processed without interrupts.
Conditioning on the first two events yields

$$\Psi_r(\lambda) = \frac{\Lambda_{r-1}}{\Lambda_{r-1}+\mu_r+\lambda}\left[\pi_r(\lambda) + (1-\pi_r(\lambda))\Psi_r(\lambda)\right] + \frac{\lambda}{\Lambda_{r-1}+\mu_r+\lambda}, \quad (6.6.8)$$

which may be rewritten as

$$\Psi_r(\lambda) = \frac{\lambda + \Lambda_{r-1}\pi_r(\lambda)}{\mu_r + \lambda + \Lambda_{r-1}\pi_r(\lambda)}. \quad (6.6.9)$$

Relations (6.6.6), (6.6.7) and (6.6.9) induce a recursive scheme for the evaluation of the probabilities $\pi_r(\lambda)$.

Before starting the discussion on general processing time distributions, an alternative formulation of the recursion is given establishing the equivalence of our scheme and the scheme proposed in Veran [1984]. It is assumed that all processing times are exponentially distributed.

Lemma 6.1

Define $\theta_r(\lambda)$, for $r=1,\dots,R$ and $\lambda \geqslant 0$, as

$$\theta_r(\lambda) = \lambda + \Lambda_{r-1}\pi_r(\lambda). \quad (6.6.10)$$

Then, for $r=1$ and $\lambda \geqslant 0$,

$$\theta_1(\lambda) = \lambda \quad (6.6.11)$$

and, for $r=2,\dots,R$ and $\lambda \geqslant 0$,

$$\theta_r(\lambda) = \theta_{r-1}(\lambda)\,\frac{\mu_{r-1} + \theta_{r-1}(\lambda+\lambda_{r-1})}{\mu_{r-1} + \theta_{r-1}(\lambda)}. \quad (6.6.12)$$

Proof

For $r=1$ and $\lambda \geqslant 0$ (6.6.11) trivially holds. So, fix r, $r=1,\dots,R-1$ and λ, $\lambda \geqslant 0$. From (6.6.5) it follows that

$$\theta_{r+1}(\lambda) = \lambda + \lambda_r\Psi_r(\lambda) + \Lambda_{r-1}\Phi_r(\lambda). \quad (6.6.13)$$

Furthermore, it follows from (6.6.9) that

$$\Psi_r(\lambda) = \frac{\theta_r(\lambda)}{\mu_r + \theta_r(\lambda)} \quad (6.6.14)$$

and from (6.6.7) and (6.6.14) that

$$\Lambda_{r-1}\Phi_r(\lambda) = \theta_r(\lambda)-\lambda+\frac{[\theta_r(\lambda+\lambda_r)-\lambda_r-\theta_r(\lambda)]\,\theta_r(\lambda)}{\mu_r+\theta_r(\lambda)}\,. \tag{6.6.15}$$

Combining (6.6.13), (6.6.14) and (6.6.15) yields (6.6.12). ∎

The Relations (6.6.11) and (6.6.12) correspond with the scheme as given by Relation (13) in Veran [1983].

General processing time distributions

Let G_r be the processing time distribution of a job from a particular terminal r. Conditioning on the processing time of the job from terminal r we find for $\lambda \geqslant 0$

$$\Psi_r(\lambda) = \int_0^\infty p_r(\lambda,x)\,dG_r(x)\,, \tag{6.6.16}$$

where $p_r(\lambda,x)$ is the conditional probability that during a busy r-period at least one customer of a Poisson process with parameter λ arrives, provided this busy r-period starts with a job from terminal r with a processing time x.

The processing time of a job from terminal r at the central processor is interrupted by jobs of higher priority forming busy (r-1)-periods. Conditioning on the number of busy (r-1)-periods interrupting the processing time of a job from terminal r of length x, we find

$$p_r(\lambda,x) = \sum_{k=0}^{\infty} \frac{(\Lambda_{r-1}x)^k}{k!}\, e^{-\Lambda_{r-1}x}\; p_r(\lambda,x,k)\,, \tag{6.6.17}$$

where $p_r(\lambda,x,k)$ is the conditional probability that during a busy r-period at least one customer from a Poisson process with parameter λ arrives, provided this busy r-period starts with a job from terminal r with a processing time x which is interrupted by k busy (r-1)-periods.

One may verify that

$$p_r(\lambda,x,k) = 1 - e^{-\lambda x}\,(1-\pi_r(\lambda))^k\,. \tag{6.6.18}$$

Combining these results yields

$$\Psi_r(\lambda) = \int_0^\infty [1 - e^{-x(\lambda+\Lambda_{r-1}\pi_r(\lambda))}]\,dG_r(x)\,. \tag{6.6.19}$$

If we write the Laplace-Stieltjes transform of the distribution function G_r as Γ_r, i.e. for all $s \geqslant 0$

$$\Gamma_r(s) = \int_0^\infty e^{-sx}\,dG_r(x)\,, \tag{6.6.20}$$

then $\Psi_r(\lambda)$ may be written as

$$\Psi_r(\lambda) = 1 - \Gamma_r(\lambda + \Lambda_{r-1}\pi_r(\lambda)) . \tag{6.6.21}$$

This completes the analysis of $\Psi_r(\lambda)$ for a general processing time distribution.

The last relation is interesting because it shows that for processing times with a rational Laplace-Stieltjes transform, the evaluation of $\Psi_r(\lambda)$ reduces to the solution of a set of linear equations, cf. Van Doremalen [1984:2] for a discussion on phase-type processing time distributions.

6.6.3. A CP terminal system with multi-terminal groups

The analysis discussed in the previous section can be extended to certain CP-terminal systems with more than one terminal per group. Such multi-terminal groups consist of sets of active terminals with identical specifications. The think times of the users and the processing times of the jobs are stochastically independent and exponentially distributed. The priority level is the same for all terminals in a specific group. The service at the CP is governed by an alternative service discipline which is assumed to be work-conserving, i.e. it does not affect the amount of processing time of a given customer, cf. Wolff [1970].

The basic idea in the analysis of a system with multi-terminal groups is a conversion of the service discipline at the CP. Each terminal obtains its own priority level. So, there are as many priority levels as terminals. The resulting model is a CP terminal model which may be analysed with the technique that has been described in the previous section. An aggregation step provides the results for the original system with multi-terminal groups.

Essential in the justification of the proposed analysis is to determine whether the utilization of the CP is influenced by the transformation of the service discipline. In general, the service discipline will have an influence on the utilization.
However, if the processing times at the CP are exponentially distributed, the utilization is not affected by applying an alternative work-conserving service discipline. Examples of work-conserving disciplines are first-come first-served, processor sharing and, for our analysis very important, preemptive resume priorities.
To understand that the utilization is not affected by the use of an alternative work-conserving service discipline, note that the jobs from terminals with the same exponentially distributed processing times are in a sense undistinghuishable and therefore interchangeable at the CP.

This subsection is concluded with an elaboration of this idea for the evaluation of the utilizations in a multi-terminal CP-terminal system.

Let the tuple (r,k) denote terminal number k in group r, where $r = 1,\ldots,R$ and, for given r, $k = 1,\ldots,K_r$. The auxiliary priority rule is introduced as follows. A job from terminal (i,l) has a higher priority than a job from terminal (r,k) if $i < r$ and, if i=r, if $l < k$.
Invoking the recursive scheme derived in the preceding subsection one may evaluate the

fractions $u_{r,k}$ of time the CP is executing jobs from the various terminals (r,k). An aggregation step yields the fraction u_r of time the CP is executing jobs from group r

$$u_r = \sum_{k=1}^{K_r} u_{r,k} \ . \tag{6.6.22}$$

The expected response time of a job from terminal group r is evaluated by

$$S_r = \frac{K_r w_r}{u_r} - \frac{1}{\lambda_r} \ . \tag{6.6.23}$$

6.6.4. Conclusions and remarks

Based on mean value ideas and renewal arguments a recursive scheme has been derived for the evaluation of important characteristics of a system comprising a set of terminals and a central processor with a preemptive resume priority discipline. It should be observed that the evaluation of the recursive scheme is not trivial. Of course, it is possible to implement the recursion by brute force in any high level programming language with a recursion feature, like in ALGOL 68. For larger values of $R, K_1, \ldots, K_R$ this cannot be recommended, as the computational complexity and storage requirements of such an implementation would be disastrous. In Van Doremalen [1984:3] we have presented an enumeration algorithm which is relatively efficient. The number of recursion steps is in the order of

$$\left(\sum_{r=1}^{R} K_r\right) \prod_{r=1}^{R} (K_r + 1) \ .$$

The complexity of the scheme thus resembles that of the MVA algorithm for the evaluation of a closed multichain queueing network model with R chains and population vector $K_1, \ldots, K_R$. An immediate consequence is that for larger numbers of priority levels an exact analysis is prohibited by the complexity of the recursion. The standard approximation methods of Section 6.4 have the same complexity as the exact method introduced in this section. For larger values of R one has, therefore, to resort to the methods introduced in Section 6.5. In the next section some examples are presented which support these remarks.

6.7. Numerical results and conclusions

6.7.1. Introduction

In this section we discuss numerical results illustrating the accuracy of the various methods presented for two typical models from the analysis of computer systems: a CP-terminal system and a closed central server model.

We have restricted the discussion to a few examples which represent a large class of queueing network models. For a more detailed numerical analysis of the methods and a

comparison with other approximation methods we refer to Van Doremalen, Wessels and Wijbrands [1985].

In Subsection 6.7.2 some examples of a CP-terminal system are treated. In Subsection 6.7.3 a large example is presented which is related with the closed central server model which has been introduced in Subsection 4.6.4. Some general conclusions that may be drawn from a more sophisticated analysis of the methods are reviewed in Subsection 6.7.4.

6.7.2. CP-terminal models.

The CP-terminal model provides an interesting class of test examples. In the first place because the model has its virtue as a tool for the analysis of the influence of priority schedules on the performance of computer systems. In the second place because the model can be analysed exactly in a relatively efficient way as we have shown in Section 6.6.

Our numerical example concerns four closely related models, numbered 1 till 4. A model is given by the number of groups, by the numbers of terminals per group, the expected think times and the distribution functions of the processing times. Table 6.1 shows the relevant model parameters.

Let us first consider the case where all processing times are exponentially distributed. Tables 6.2 and 6.3 show the numerical results for utilizations at the CP, i.e. the fractions of time the central processor is processing jobs of a certain group, and the expected response times at the CP per terminal group.
Apart from the exact results (EXACT) we have included the figures for the six approximation methods which have been suggested in this chapter.
The first three methods are the MVA approximations. Apart from the standard MVA method we have implemented the first order depth improvement of the Schweitzer method (SW-DI) and of the global aggregation method (AG-DI). The next three methods are the corresponding service completion time or SCT approximations.
The columns M and G stand for the model and terminal group number respectively.

The first observation that may be drawn from the tables are that the SCT approximations have a higher accuracy than the MVA approximations and that the first order depth improvements of the Schweitzer and global aggregation method yield approximately the same results as the corresponding standard methods.
These observations may be generalized as further numerical evidence has learned. The first conclusion is therefore to use the SCT approximation with a first order depth improvement of the Schweitzer or global aggregation method.

The second observation is that the rather artificial priority schedules of the models 2 and 4 yield relatively bad results, whereas the more natural schedules which give preference to the smaller jobs yield very good results. This observation has been made earlier for related approximation methods as for instance the virtual server approximation, cf.

Sevcik [1978] and Kaufman [1984], but numerical experiments have indicated that it is true for most approximation methods presented in the literature, cf. Van Doremalen, Wessels and Wijbrands [1985] for a discussion.
The second conclusion is therefore to be careful when studying priority schedules which give preference to jobs with longer processing times.

Apart from this CP-terminal system we present the results for the models 1 and 2 with deterministic service demands at the CP in Table 6.4. We have used the first order depth improvement of the Schweitzer method. The results are very remarkable if we compare the bad approximations for model 2 with exponential processing times with the relatively good approximations for the model with deterministic processing times. Further research in this direction is necessary to explain this behaviour.

M	populations				processing times				think times			
G	1	2	3	4	1	2	3	4	1	2	3	4
1	1	1	2	2	1	16	1	16	4	64	8	128
2	1	1	2	2	2	8	2	8	8	32	16	64
3	1	1	2	2	4	4	4	4	16	16	32	32
4	1	1	2	2	8	2	8	2	32	8	64	16
5	1	1	2	2	16	1	16	1	64	4	128	8

Table 6.1 : Model parameters for the CP-terminal examples.

			MVA approximations			SCT approximations		
M	G	EXACT	MVA	SW-DI	AG-DI	SCT	SW-DI	AG-DI
1	1	.200	.200	.200	.200	.200	.200	.200
	2	.187	.186	.186	.186	.187	.187	.187
	3	.172	.169	.169	.169	.171	.171	.172
	4	.151	.147	.147	.148	.151	.151	.152
	5	.124	.118	.118	.119	.124	.124	.125
2	1	.200	.200	.200	.200	.200	.200	.200
	2	.181	.174	.174	.172	.177	.177	.176
	3	.153	.127	.127	.125	.140	.140	.138
	4	.122	.073	.073	.072	.095	.095	.093
	5	.093	.036	.036	.034	.055	.055	.055
3	1	.220	.220	.220	.219	.220	.220	.219
	2	.208	.208	.208	.207	.208	.208	.208
	3	.192	.190	.190	.190	.193	.193	.193
	4	.167	.162	.162	.163	.168	.168	.168
	5	.126	.118	.118	.120	.126	.127	.128
4	1	.220	.220	.220	.219	.220	.220	.219
	2	.202	.197	.197	.196	.200	.200	.199
	3	.171	.147	.148	.147	.163	.163	.161
	4	.132	.083	.083	.083	.111	.110	.110
	5	.094	.039	.039	.038	.063	.063	.063

Table 6.2: Approximations of the utilizations at the CP.

M	G	EXACT	MVA approximations			SCT approximations		
			MVA	SW-DI	AG-DI	SCT	SW-DI	AG-DI
1	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	2	2.7	2.8	2.8	2.8	2.7	2.7	2.7
	3	7.3	7.7	7.7	7.6	7.4	7.4	7.3
	4	20.9	22.5	22.5	22.1	21.0	21.0	20.6
	5	64.9	71.5	71.5	70.7	65.1	65.1	63.8
2	1	16.0	16.0	16.0	16.0	16.0	16.0	16.0
	2	12.3	14.0	14.0	14.4	13.2	13.2	13.5
	3	10.1	15.4	15.4	15.9	12.5	12.5	12.9
	4	8.4	19.2	19.2	19.9	13.1	13.1	13.4
	5	6.8	23.5	23.5	25.2	14.1	14.1	14.2
3	1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	2	3.2	3.3	3.3	3.3	3.2	3.2	3.2
	3	9.6	10.2	10.1	10.1	9.5	9.5	9.6
	4	31.8	35.3	34.7	34.2	31.4	31.3	31.0
	5	126.2	144.0	142.8	139.7	125.8	124.1	121.1
4	1	17.8	17.8	17.8	17.9	17.8	17.8	17.9
	2	15.3	17.2	17.3	17.7	16.0	16.0	16.4
	3	14.7	22.2	22.2	22.5	17.1	17.2	17.6
	4	14.2	31.8	31.9	32.4	20.1	20.2	20.5
	5	13.2	42.7	42.9	44.4	23.8	23.9	23.9

Table 6.3 : Approximations for the expected response times at the CP.

M	G	deterministic processing times		exponential processing times	
		EXACT	SCT-SW-DI	EXACT	SCT-SW-DI
1	1	.200	.200	.200	.200
	2	.189	.189	.187	.187
	3	.174	.174	.172	.171
	4	.154	.154	.151	.151
	5	.127	.125	.124	.124
2	1	.200	.200	.200	.200
	2	.184	.180	.181	.177
	3	.159	.159	.153	.140
	4	.127	.122	.122	.095
	5	.095	.094	.093	.055

Table 6.4 : Approximations for the utilizations at the CP for deterministic and exponential processing time distributions.

6.7.3. A closed central server model

The closed central server model introduced in Subsection 4.6.4 provides an interesting example for the approximate analysis of large scale computer systems with a preemptive resume priority schedule at the central processor units.

The system comprises three CPU's and nine I/O devices. Three types of jobs are distributed over the three CPU's to form nine closed customer chains in a closed multichain queueing network model. Let us assume that one wants to study the influence of a priority schedule at the CPU's. The idea is to give preference to type 1 over type 2 and to type 2 over type 3.
Table 6.5 pictures the throughputs of the three types at the respective CPU's. Apart from the results obtained by a simulation procedure (SIMULATION) where the 95 % confidence region is indicated (CONFIDENCE), the results of the first order depth improvements of the Schweitzer and global aggregation method have been evaluated for the MVA and SCT approximations.

Observe that the results support the conclusions drawn in the preceding subsection. The figures should also be compared with these in Table 4.13, where the results for the equivalent system without a priority schedule are given.

	type 1			type 2			type 3		
CPU	1	2	3	1	2	3	1	2	3
SIMULATION	34.6	11.3	11.4	13.2	13.8	7.1	3.8	7.2	9.4
CONFIDENCE	0.4	0.2	0.2	0.4	0.4	0.2	0.2	0.2	0.3
MVA-SW-DI	34.6	11.2	11.4	13.2	13.6	6.9	3.8	6.7	9.3
MVA-AG-DI	35.1	11.4	11.6	13.5	13.9	7.0	3.9	6.8	9.2
SCT-SW-DI	34.5	11.2	11.4	13.2	13.6	6.9	3.8	6.7	9.3
SCT-AG-DI	35.0	11.4	11.6	13.6	13.8	7.0	4.2	7.2	9.4

Table 6.5 : Approximations for the throughputs of the different types at the different central processor units.

6.7.4. Conclusions

We have presented a new and promising approximation method for the analysis of closed queueing network models with a preemptive resume priority schedule at some of the resources. Regrettably, the computational complexity prohibits an evaluation of the methods for larger numbers of priority levels. This problem has been solved in an efficient way be the application of the approximation methods for queueing network models with many closed customer chains that have been developed in Chapter 4. As the method is an extension of the standard MVA algorithm the implementation in existing software packages is straightforward.

In Van Doremalen, Wijbrands and Wessels [1985] we have presented a comparison of a number of approximation methods that have been proposed in the literature. It has appeared that the SCT or service completion time approximation provides a reasonable level of accuracy at low computational costs, especially if the first order depth improvements of the Schweitzer and global aggregation method are applied. This makes the method an interesting tool for the performance evaluation of systems with preemptive resume priority queues.

It should be noted that the method allows for a number of extensions.
In Wijbrands [1985] the special layout of the closed central server model is exploited to design an improved version of the SCT approximation. This model has an important albeit limited field of application.
Extensions of the suggested ideas to mixed open and closed queueing network models and models with a head-of-the-line priority schedule are possible and relatively straightforward. The first numerical experiments in that direction are promising, but it is premature to draw to strong a conclusion.

The analysis presented and the numerical results support one of the major outcomes of our research. For the analysis of closed queueing network models the use of strictly recursive approximation methods yield attractive intuitive, efficient and accurate approximation methods.

References

AVI-ITZHAK, B. AND HEYMAN, D.P., "Approximate queueing models for multiprogramming computer systems," *O.R.*, vol. 21, pp. 1212-1230, 1973.

BARD, Y., "Some extensions to multiclass queueing network analysis," in *Fourth International Symposium on Modelling and Performance Evaluation of Computer Systems*, ed. M. Arato, A. Butrimenko, E. Gelenbe, North Holland, Amsterdam, 1979.

BASKETT, F., CHANDY, K.M., MUNTZ, R., AND PALACIOS-GOMEZ, F., "Open, closed and mixed networks of queues with different classes of customers," *J.A.C.M.*, vol. 22, pp. 248-260, 1975.

BERMON, A. AND PLEMMONS, R.J., *Nonnegative matrices in mathematical sciences*, Academic Press, New York, 1979.

BONDI, A., "Modelling the effect of local area network contention on the performance of host computers," in *Performance '84*, ed. E. Gelenbe, North-Holland, Amsterdam, 1984.

BRATLEY, P., FOX, B.L., AND SCHRAGE, L.E., *A guide to simulation*, Springer Verlag, Berlin, 1983.

BRUELL, S.C. AND BALBO, G., *Computational algorithms for closed queueing networks*, North-Holland, Amsterdam, 1980.

BRUELL, S.C., BALBO, G., AND AFSHARI, P.V., "Mean Value Analysis of mixed multiple class BCMP-networks with load dependent service stations," *Perf. Eval.*, vol. 4, pp. 241-260, 1984.

BRUMELLE, S.L., "On the relation between customer and time averages in queues," *J.A.P.*, vol. 8, pp. 508-520, 1971.

BRYANT, R., KRZESINSKI, A., AND TEUNISSEN, P., "The MVA pre-empt resume priority approximation," *Perf. Eval. Rev.*, vol. 5, pp. 12-27, 1983.

BUZEN, J., "Computational algorithms for closed queueing networks with exponential servers," *Comm. of the A.C.M.*, vol. 16, pp. 527-531, 1973.

CHANDY, K.M., HERZOG, U., AND WOO, L., "Parametric analysis of queueing networks," *I.B.M. Journal of Res. and Dev.*, vol. 19, pp. 36-42, 1975.

CHANDY, K.M., HERZOG, U., AND WOO, L., "Approximate analysis of general queueing networks," *I.B.M. Journal of Res. and Dev.*, vol. 19, pp. 43-49, 1975.

CHANDY, K.M., HOWARD, J., AND TOWSLEY, D., "Product form and local balance in queueing networks," *Journal of the A.C.M.*, vol. 24, pp. 250-263, 1977.

CHANDY, K.M. AND SAUER, C.H., "Approximate methods for analyzing queueing network models of computing systems," *Comp. Surv.*, vol. 10, pp. 281-317, 1978.

CHANDY, K.M. AND SAUER, C.H., "Computational algorithms for product form queueing networks," *Comm. of the A.C.M.*, vol. 23, pp. 573-583, 1980.

CHANDY, K.M. AND NEUSE, D., "Linearizer: A heuristic algorithm for queueing network models of computing systems," *Comm. of the A.C.M.*, vol. 25, pp. 126-134, 1982.

CHANDY, K.M. AND LAKSHMI, M.S., "An approximation technique for queueing networks with preemptive priority queues," Report, University of Austin, Dept. of Comp. Sci., Austin, Texas, 1983.

CHOW, W.M., "Approximations for large scale closed queueing networks," *Perf. Eval.*, vol. 3, pp. 1-12, 1983.

CHOW, W.M. AND YU, P.S., "An approximation technique for central server queueing models with a priority dispatching rule," *Per. Eval.*, vol. 3, pp. 55-62, 1983.

CHUNG, K.L., *Markov chains with stationary transition probabilities*, Springer Verlag, Berlin, 1960.

CINLAR, E., *Introduction to stochastic processes*, Prentice Hall, Englewood Cliffs, N.J., 1975.

COBHAM, A., "Priority assignment in waiting line problems," *O.R.*, vol. 2, pp. 70-76, 1954.

CONWAY, A.E. AND GEORGANAS, N.D., "RECAL: A new efficient algorithm for the exact analysis of multiple chain closed queueing networks," Rapport de Recherche No 373, I.N.R.I.A., Le Chesnay Cedex, France, 1985.

COURTOIS, P.J., *Decomposability: queueing and computer system applications*, Academic Press, New York, 1977.

DOREMALEN, J.B.M. VAN, "Mean Value Analysis in multichain queueing networks: an iterative approximation," in *Operations research proceedings 1983*, ed. H. Steckhan, W. Buhler, K.E. Jager, Ch. Schneeweiss, J. Schwarze, Springer Verlag, Berlin, 1984.

DOREMALEN, J.B.M. VAN, "A Mean Value Analysis of a closed CP-terminal system with preemptive resume priorities," in *Performance '84*, ed. E. Gelenbe, North-Holland, Amsterdam, 1984.

DOREMALEN, J.B.M. VAN, "An algorithm for the evaluation of performance measures in a CP-terminal system," Memorandum COSOR 84-07, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., Eindhoven, 1984.

DOREMALEN, J.B.M. VAN AND WAAL, P.R. DE, "An approximation method for closed queueing networks with two-phase servers," Memorandum COSOR 85-15, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., Eindhoven, 1985.

DOREMALEN, J.B.M. VAN AND WESSELS, J., "Iterative approximations for networks of queues," in *Stochastic Programming*, ed. F. Archetti, G. di Pillo, M. Lucertini, Springer Verlag, Berlin, 1985.

DOREMALEN, J.B.M. VAN, WESSELS, J., AND WIJBRANDS, R., "Approximate analysis of priority queueing networks," in *Proceedings of the International Seminar on Teletraffic Analysis and Computer Performance Evaluation*, North-Holland, Amsterdam, 1985.

EAGER, D.L. AND SEVCIK, K.C., "Performance bound hierarchies for queueing networks," *A.C.M. Trans. on Comp. Syst.*, vol. 1, pp. 99-115, 1983.

EAGER, D.L. AND SEVCIK, K.C., "An analysis of an approximation algorithm for queueing networks," *Perf. Eval.*, vol. 4, pp. 275-284, 1984.

FERRARI, D., *Computer systems performance evaluation*, Prentice Hall, Englewood Cliffs, N.J., 1978.

GAVER, D.P., "A waiting line with interrupted service including priorities," *Journal of the Royal Statistical Society*, vol. B 25, pp. 73-90, 1962.

GELENBE, E. AND MITRANI, I., *Analysis and synthesis of computer systems*, Academic Press, New York, 1980.

GORDON, W.J. AND NEWELL, G.F., "Closed queueing systems with exponential servers," *O.R.*, vol. 15, pp. 254-265, 1967.

HEIDELBERGER, P. AND LAVENBERG, S.S., "Computer performance evaluation methodology," *I.E.E.E. Transactions on Computers*, vol. 33, pp. 1195-1216, 1985.

HEYMAN, D. AND STIDHAM, S., "The relation between customer and time averages in queues," *O.R.*, vol. 28, pp. 983-994, 1980.

HEYMAN, D. AND SOBEL, M., *Stochastic models and operations research: volume I*, McGraw-Hill Book Company, New York, 1982.

HINE, J.H., MITRANI, I., AND TSUR, S., "The control of response times in multiclass systems by memory allocation," *Comm. of the A.C.M.*, vol. 22, pp. 415-424, 1979.

HOORN, M. VAN, "Algorithms and approximations for queueing systems," Thesis, Free University of Amsterdam, Amsterdam, 1983.

HORDIJK, A. AND DIJK, N. VAN, "Stationary probabilities for networks of queues," Report 81-19, University of Leiden, Inst. for Applied Math. and Comp. Sci., Leiden, 1981.

HORDIJK, A. AND DIJK, N. VAN, "Networks of queues: Parts I and II," in *Modelling and Performance Evaluation Methodology*, ed. F. Bacelli, G. Fayolle, Springer Verlag, Berlin, 1984.

JACKSON, J.R., "Networks of waiting lines," *O.R.*, vol. 5, pp. 518-521, 1957.

JAISWAL, N.K., *Priority Queues*, Academic Press, New York, 1968.

KAMEDA, H., "Optimality of a central processor scheduling policy for processing a job stream," *A.C.M. Trans. on Comp. Syst.*, vol. 2, pp. 78-90, 1984.

KAUFMAN, J.S., "Approximation methods for networks of queues with priorities," *Perf. Eval.*, vol. 4, pp. 183-198, 1984.

KELLY, F., "Networks of queues," *A.A.P.*, vol. 8, pp. 416-432, 1976.

KELLY. F., *Reversibility and stochastic networks*, John Wiley and Sons, New York, 1979.

KLEINROCK, L., *Queueing Systems, Volume 1: Theory*, John Wiley and Sons, New York, 1975.

KLEINROCK, L., *Queueing Systems, Volume 2: Computer Applications*, John Wiley and Sons, New York, 1975.

KOBAYASHI, H., *Modelling and analysis: an introduction to system performance evaluation methodology*, Addison-Wesley, Reading, Mass., 1978.

KRZESINSKI, A., TEUNISSEN, P., AND KRITZINGER, P., "Mean Value Analysis for load dependent servers in mixed multiclass queueing networks," ITR 82-01-00, University of Stellenbosch, Inst. for Appl. Comp. Sci., Stellenbosch, South Africa, 1982.

KUEHN, P.J., "Approximate analysis of general queueing networks by decomposition," *I.E.E.E. Trans. on Comm.*, vol. COM 27, pp. 113-126, 1979.

LAM, S.S., "Queueing networks with population size constraints," *I.B.M. Journal of Res. and Dev.*, vol. 21, pp. 370-378, 1977.

LAM, S.S. AND WONG, J.W., "Queueing network models of packet switching networks. Part 2: Networks with population size constraints," *Perf. Eval.*, vol. 2, pp. 161-180, 1982.

LAM, S.S. AND LIEN, Y.L., "A tree convolution algorithm for the solution of queueing networks," *Comm. of the A.C.M.*, vol. 26, pp. 203-215, 1983.

LAVENBERG, S.S. AND REISER, M., "Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers," *J.A.P.*, vol. 17, pp. 1048-1061, 1980.

LAVENBERG, S.S., *Computer performance modelling handbook*, Academic Press, New York, 1983.

LAVENBERG, S.S. AND SAUER, C.H., "Approximate analysis of queueing networks," in *Computer performance modelling handbook*, ed. S.S. Lavenberg, Academic Press, 1983.

LAZAR, A.A. AND ROBERTAZZI, T.G., *The geometry of lattices for Markovian queueing networks*, Preprint, 1984.

LAZOWSKA, E.D., ZAHORJAN, J., GRAHAM, G.S., AND SEVCIK, K.C., *Quantitive system performance*, Prentice Hall, Englewood Cliffs, N.J., 1984.

LITTLE, J.D., "A proof of the queueing formula L=λW," *O.R.*, vol. 22, pp. 417-421, 1961.

MARIE, R. AND STEWART, W.J., "A hybrid iterative-numerical method for the solution of a general queueing network," in *Measuring, modelling and evaluating computer systems*, ed. H. Beilner, E. Gelenbe, North-Holland, Amsterdam, 1977.

MARIE, R., "An approximate analytical method for general queueing networks," *I.E.E.E. Trans. on Software Engineering*, vol. SE-5, pp. 530-538, 1979.

MARIE, R., "Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queues," *A.C.M. Perf. Eval. Rev.*, pp. 117-125, 1980.

MARKOWITZ, H.M., "Simulator design and programming," in *Computer performance modelling handbook*, ed. S.S. Lavenberg, Academic Press, New York, 1983.

MARKS, B.I., "State probabilities of M/M/1 priority queues," *O.R.*, vol. 21, pp. 974-987, 1973.

MORRIS, R.J.T., "Priority queueing networks," *Bell Systems Technical Journal*, vol. 60, pp. 1745-1769, 1981.

NEUSE, D. AND CHANDY, K.M., "HAM: the heuristic aggregation method," *A.C.M. Perf. Eval. Rev.*, vol. 5, pp. 195-212, 1983.

NEUTS, M.F., *Matrix-geometric solutions in stochastic models: an algorithmic approach*, John Hopkins University Press, Baltimore, 1981.

OLIVER, R.M., "An alternative derivation of the Pollaczek-Khintchine formula," *O.R.*, vol. 12, pp. 158-159, 1964.

ORTEGA, J.M. AND RHEINBOLDT, W.C., *Iterative solutions of non-linear equations in several variables*, Academic Press, New York, 1970.

REISER, M. AND KOBAYASHI, H., "Queueing networks with multiple closed chains: theory and computational algorithms," *I.B.M. Journal of Res. and Dev.*, vol. 19, pp. 283-294, 1975.

REISER, M., "Interactive modelling of computer systems," *I.B.M. Systems Journal*, vol. 16, pp. 309-327, 1976.

REISER, M., "Mean Value Analysis of queueing networks: a new look at an old problem," in *Fourth Int. Symp. on Modelling and Performance Evaluation of Computer Systems*, ed. M. Arato, A. Butrimenko, E. Gelenbe, North-Holland, Amsterdam, 1979.

REISER, M., "A queueing network analysis of computer communication networks with window flow control," *I.E.E.E. Trans. on Comm.*, vol. COM-27, pp. 1199-1209, 1979.

REISER, M. AND LAVENBERG, S.S., "Mean Value Analysis of closed multichain queueing networks," *Journal of the A.C.M.*, vol. 27, pp. 313-322, 1980.

REISER, M., "Mean Value Analysis and convolution method for queue-dependent servers in closed queueing networks," *Perf. Eval.*, vol. 1, pp. 7-18, 1981.

ROBERTAZZI, T.G. AND LAZAR, A.A., "On the modelling and optimal flow control of the Jacksonian network," *Perf. Eval.*, vol. 5, pp. 29-43, 1985.

ROSS, S., *Stochastic Processes*, Prentice Hall, Englewood Cliffs, N.J., 1983.

SAUER, C.H. AND CHANDY, K.M., "Approximate analysis of central server models," *I.B.M. Journal of Res. and Dev.*, vol. 19, pp. 301-313, 1975.

SAUER, C.H. AND CHANDY, K.M., *Computer systems performance modelling*, Prentice Hall, Englewood Cliffs, N.J., 1981.

SCHASSBERGER, R., "An aggregation principle for computing invariant probability vectors of semi-Markovian models," in *Mathematical computer performance and reliability*, ed. G. Iazeolla, P.J. Courtois, A. Hordijk, North-Holland, Amsterdam, 1984.

SCHMITT, W., "Approximate analysis of Markovian queueing networks with priorities," in *Performance '84*, ed. E. Gelenbe, North-Holland, Amsterdam, 1984.

SCHWEITZER, P.J., "Approximate analysis of multiclass closed networks of queues," in *International Conference on Stochastic Control and Optimization*, Amsterdam, 1979.

SCHWEITZER, P.J., "Aggregation methods for large Markov chains," in *Mathematical computer performance and reliability*, ed. G. Iazeolla, P.J. Courtois, A. Hordijk, Nort-Holland, Amsterdam, 1984.

SENETA, E., *Non-negative matrices: an introduction to theory and applications*, George Allen and Unwin Ltd., London, 1973.

SEVCIK, K.C., "Priority scheduling disciplines in queueing network models of computer systems," in *Proceedings of the IFIP-congress '77*, ed. B. Gilchrist, North-Holland, Amsterdam, 1978.

SEVCIK, K.C. AND MITRANI, I., "The distribution of queueing network states at input and output instants," *Journal of the A.C.M.*, vol. 28, pp. 358-371, 1981.

SILVA, E. DE SOUZA E, LAVENBERG, S.S., AND MUNTZ, R., "A perspective on iterative methods for the approximate analysis of closed queueing networks," in *Mathematical computer performance and reliability*, ed. G. Iazeolla, P.J. Courtois, A. Hordijk, North-Holland, Amsterdam, 1984.

STIDHAM, S., "Regenerative processes in the theory of queues with applications to the alternating priority queue," *A.A.P.*, vol. 4, pp. 542-577, 1972.

STIDHAM, S., "A last word on $L=\lambda W$," *O.R.*, vol. 22, pp. 417-421, 1974.

SURI, R., "A concept of monotonicity and its characterization for closed queueing networks," *O.R.*, vol. 33, pp. 606-624, 1985.

TAKACS, L., "Priority queues," *O.R.*, vol. 12, pp. 63-74, 1964.

TIJMS, H. AND HOORN, M. VAN, "Algorithms for the state probabilities and waiting times in single server queueing systems with random and quasi random input and phase-type service times," *OR Spektrum*, vol. 2, pp. 145-152, 1981.

TUCCI, S. AND SAUER, C.S., "The Tree MVA-algorithm," *Perf. Eval.*, vol. 5, pp. 187-196, 1985.

VANTILBORGH, H., "Exact aggregation in exponential queueing networks," *Journal of the A.C.M.*, vol. 25, pp. 620-629, 1978.

VERAN, M., "Exact analysis of a priority queue with finite sources," in *Modelling and performance evaluation methodology*, ed. F. Bacelli, G. Fayolle, Springer Verlag, Berlin, 1984.

WAL, J. VAN DER, "CP-utilization in an exponential CP-terminal system with equal think times and different job sizes," Memorandum COSOR 82-13, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., Eindhoven, 1982.

WAL, J. VAN DER, "Monotonicity of the throughput of a closed exponential queueing network in the number of jobs," Memorandum COSOR 85-21, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., Eindhoven, 1985.

WELCH, P.D., "On a generalized M/G/1 queueing process in which the first customer of each busy period receives exceptional service," *O.R.*, vol. 12, pp. 736-752, 1964.

WIJBRANDS, R., "On an approximation method for priority queueing in CPU-disk models," Memorandum COSOR 85-18, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., Eindhoven, 1985.

WOLFF, R.W., "Work conserving priorities," *J.A.P.*, vol. 7, pp. 327-337, 1970.

WOLFF, R.W., "Poisson arrivals see time averages," *O.R.*, vol. 30, pp. 223-231, 1982.

ZAHORJAN, J. AND WONG, E., "The solution of separable queueing networks using Mean Value Analysis," *A.C.M. Sigm. Perf. Eval. Rev.*, vol. 3, pp. 80-85, 1981.

Samenvatting

De laatste twee decennia is er zowel van theoretische als praktische zijde een groeiende belangstelling voor de analyse van grote en complexe wachtrijsystemen.
Aan de theoretische zijde heeft de snelle ontwikkeling van de moderne rekenapparatuur voor een doorbraak gezorgd in de bestudering van algoritmisch georienteerde technieken voor de analyse van min of meer complexe wiskundige modellen van netwerken van wachtrijen.
Aan de praktische zijde groeit de behoefte aan efficiente en betrouwbare hulpmiddelen voor de ontwikkeling, bestudering en verbetering van bijvoorbeeld computer systemen, telecommunicatie netwerken, produktielijnen en transport netwerken.
Het gebruik van wachtrijmodellen wordt op een natuurlijke manier gerechtvaardigd door een gemeenschappelijke karakteristiek van dergelijke systemen. Ze kunnen worden opgevat als een verzameling van onderling verbonden werkstations, die taken verrichten voor verschillende groepen klanten. Omdat werksnelheden en kapaciteiten van de werkstations in het algemeen eindig zijn onstaan voor elk van de werkstations wachtrijen.

In de praktijk is simulatie nog altijd het belangrijkste gereedschap bij de analyse van wachtrijsystemen. Twee nauw samenhangende recente ontwikkelingen kunnen echter een verschuiving teweeg brengen in de richting van het gebruik van analytische technieken. Voor een interessante deelklasse van netwerken zijn namelijk efficiente algoritmen ontwikkeld om belangrijke systeemgrootheden te berekenen. Deze algoritmen vormen bovendien een veelbelovende basis voor het ontwerpen van efficiente en betrouwbare analytische benaderingsmethoden.

Bij de bestudering van wachtrijsystemen speelt de analyse van stochastische modellen en met name van kontinue-tijd Markov-processen met een diskrete toestandsparameter een belangrijke rol. Hoewel het theoretisch mogelijk is om het tijdsafhankelijke gedrag van dergelijke processen te analyseren, ligt de nadruk op de bestudering van het limietgedrag. Onder niet al te restriktieve voorwaarden is de limietverdeling namelijk de unieke, strikt positieve en genormeerde oplossing van een eindig of aftelbaar stelsel lineaire vergelijkingen, de evenwichtsvergelijkingen.
Als deze oplossing een aantrekkelijke analytische vorm heeft of op een efficiente manier kan worden berekend, is het gebruik van analytische technieken in het algemeen te prevaleren boven het gebruik van simulatie.

Een klasse van dergelijke wachtrijmodellen wordt gevormd door de separabele of produktvorm netwerken, die gekenschetst worden door het feit dat oplossing van het stelsel evenwichtsvergelijkingen een produktvorm aanneemt.

Het belang van separabele netwerken voor de analyse van wachtrijsystemen is vooral gelegen in het feit dat voor de evaluatie van belangrijke systeemgrootheden, zoals gemiddelde verblijftijden, doorstroomsnelheden en gemiddelde aantallen klanten, relatief efficiente algoritmen zijn ontwikkeld. De twee voornaamste rekentechnieken zijn het convolutie algoritme en de gemiddelde waarden analyse.

Helaas zijn de meeste realistische modellen van wachtrijsystemen niet separabel. Bovendien geldt voor grotere separabele netwerken dat de complexiteit van de resulterende rekenprocedures in vele gevallen toch een efficiente berekening van systeemgrootheden verhindert.
Het ontwikkelen van analytische benaderingsmethoden met behulp van de theorie van de separabele netwerken lijkt een natuurlijke weg om deze twee problemen aan te pakken.
In dit proefschrift wordt de aandacht met name gericht op het gebruik van de gemiddelde waarden analyse, omdat deze analyse een interessante interpretatie geeft aan een stel relaties tussen de systeemgrootheden. Hoewel deze interpretatie niet zonder meer over te zetten is op niet-separabele wachtrijmodellen, biedt ze een basis voor de ontwikkeling van heuristische benaderingsmethoden.

Het proefschrift bestaat uit twee delen. In het eerste deel, de Hoofdstukken 1,2 en 3, worden enkele algemene technieken besproken welke in het tweede deel, de Hoofdstukken 4,5 en 6, aan de hand van enkele voorbeelden worden geillustreerd.

In Hoofstuk 1 wordt een kort overzicht gegeven van de voor onze analyse relevante ontwikkelingen op het gebied van de analyse van netwerken van wachtrijen.

In Hoofdstuk 2 volgt een meer gedetailleerd overzicht van de technieken die kunnen bijdragen tot de ontwikkeling van efficiente en betrouwbare benaderingsmethoden. Met name wordt ingegaan op decompositie en aggregatie technieken en het gebruik van separabele netwerken en gemiddelde waarden analyse.

In Hoofdstuk 3 wordt een klasse van separabele netwerken besproken. Met name wordt ingegaan op de interpretatie van de gemiddelde waarden analyse. Enkele opmerkingen over de implementatie van het gemiddelde waarden algoritme worden gemaakt.

In Hoofstuk 4 wordt het complexiteitsprobleem onder de loep genomen aan de hand van een bekend probleem in de analyse van separabele netwerken: de evaluatie van systemen met meerdere gesloten klantenketens. We beschrijven een konstruktieve aanpak die leidt tot de ontwikkeling van efficiente benaderingsmethoden. Naast bestaande methoden bespreken we enkele veelbelovende nieuwe technieken.

In Hoofdstuk 5 wordt de verstoring van de separabiliteitsvoorwaarden besproken aan de hand van een wachtrijmodel met enkele zogenaamde twee fasen servers. Op min of meer geordende wijze wordt naar enkele benaderingsmethoden toegewerkt.

In Hoofdstuk 6 ontmoeten we een mengeling van de twee problemen bij de analyse van een wachtrijmodel met prioriteitswachtrijen. Het blijkt dat ook hier de structurele aanpak leidt tot efficiente en betrouwbare benaderingsmethoden.

De behandelde technieken zijn vrijwel alle geimplementeerd. De numerieke voorbeelden verschaffen enig inzicht in het gedrag van de benaderingsmethoden.

Curriculum vitae

De schrijver van dit proefschrift werd op 4 augustus 1955 geboren te Hedel, Gelderland. Van 1967 tot 1973 bezocht hij het Jacob Roelandts College te Boxtel. Na het behalen van het diploma Gymnasium-β studeerde hij van 1973 to 1982 wiskunde aan de Technische Hogeschool Eindhoven. In februari 1982 werd het ingenieursexamen wiskunde behaald met als hoofdrichting besliskunde. Het afstudeerprojekt, dat onder begeleiding van Prof.dr. J. Wessels plaats vond, betrof een onderwerp uit de wachttijdtheorie: de analyse van een tweedimensionaal wachtrijsysteem met overloop.
Sinds februari 1982 is de schrijver als wetenschappelijk assistent verbonden aan de onderafdeling der wiskunde en informatica van de Technische Hogeschool Eindhoven. Dit proefschrift is een weerspiegeling van het onderzoek dat de schrijver de afgelopen vier jaar onder begeleiding van Prof.dr. J. Wessels heeft verricht.

STELLINGEN

I

De eerste r elementen van de p-adische ontwikkeling van een rationaal getal q vormen de basis van de Hensel kode H(p,r,q). De recente ontwikkeling van efficiente koderings- en dekoderingsalgoritmen maakt het gebruik van Hensel kodes voor het exakt rekenen met rationale getallen aantrekkelijk, zie o.a. [1].

Hoewel in de literatuur in het algemeen grote priemgetallen p als basis voor de p-adische ontwikkeling worden gekozen, verdient de keuze p=2 vanuit praktisch oogpunt de voorkeur, omdat de resulterende algoritmen efficient zijn en minder geheugenruimte vergen.

[1] P. Kornerup en R.T. Gregory, " Mapping integers and Hensel codes onto Farey fractions ", B.I.T. 23(1983)9-20

II

Beschouw een geboorte-sterfte proces op de niet negatieve gehele getallen met geboortesnelheden $\lambda_i > 0$, $i = 0,1,\ldots$, en sterftesnelheden $\mu_i > 0$, $i = 1,2,\ldots$. Het n^{de} moment van de first passage time van toestand i naar toestand k wordt aangeduid met $m_i^k(n)$, $n \geqslant 0$ en $0 \leqslant i \leqslant k$. De rijvector $\boldsymbol{m}^k(n)$ met elementen $m_i^k(n)$, $i = 0,1,\ldots,k$ voldoet aan de volgende recursie:

$$m^k(0) = (1,\ldots,1),$$

$$m^k(n)A = n\boldsymbol{m}^k(n-1), \quad n > 0.$$

Hierin is

$$A = \begin{vmatrix} \lambda_0 & -\mu_1 & & & & \\ -\lambda_0 & \lambda_1+\mu_1 & -\mu_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\lambda_{k-2} & \lambda_{k-1}+\mu_{k-1} & -\mu_k \\ & & & -\lambda_{k-1} & \lambda_k+\mu_k \end{vmatrix}$$

Omdat de matrix A een tridiagonaal matrix is, bepaalt deze recursie een efficient algoritme voor de berekening van de momenten van de first passage times.

III

Er zijn uitdrukkingen afgeleid voor de blokkeringskansen in een wachtrijsysteem met eindige capaciteit en twee prioriteitsklassen, zie [1]. Met een aggregatieargument is een uitbreiding naar meerdere prioriteitsklassen mogelijk.

[1] A.S. Kapadia, M.F. Kazmi en A.C. Mitchell, " Analysis of a finite capacity nonpreemptive priority queue ", Computers and Operations Research, 11(1984)337-343.

IV

De Formule van Little, zie [1] and [2], wordt veel gebruikt in de wachttijdtheorie. Hoewel deze relatie een zeer algemene geldigheid heeft, blijft enige voorzichtigheid bij het koppelen van verwachte aantallen, doorstroomsnelheden en verwachte verblijftijden geboden. Met name dient men er op te letten dat deze grootheden op het zelfde deelsysteem en het zelfde type klanten betrekking hebben.
Zo wordt in [3] op pagina 415 het verwachte aantal lagere prioriteitsklanten in de wachtkamer ten onrechte gekoppeld aan de doorstroomsnelheid en de verwachte verblijftijd van een speciaal type lagere prioriteitsklanten.

[1] J.D. Little, " A proof of the queueing formula $L=\lambda W$ ", Operations Research 9(1961)383-387.

[2] S. Stidham, " A last word on $L=\lambda W$ ", Operations Research 22(1974)417-421.

[3] A.S. Kapadia, Y.K. Chiang en M.F. Kazmi, " Finite capacity priority queues with potential health applications ", Computers and Operations Research, 12(1985)411-420.

V

Een klasse van kontinue optimale besturingsproblemen kan door tijdsdiskretisatie vertaald worden in eindig-dimensionale optimale besturingsproblemen. Als in de diskretisatie tijdsafgeleiden worden vervangen door differentie-quotienten, dan valt een van de meest efficiente algoritmen voor eindig-dimensionale problemen samen met het overeenkomstige algoritme voor kontinue problemen onder gebruikmaking van de methode van Euler als integratiemethode voor de betreffende differentiaalvergelijkingen.

Een belangrijke konklusie, die uit deze konstatering kan worden getrokken, is dat een vroegtijdige diskretisatie van het kontinue besturingsprobleem dient te worden afgeraden.

VI

In [1] wordt het stelsel evenwichtsvergelijkingen van een twee-dimensionaal wachtrijsysteem met een overloopmogelijkheid in een enkele richting opgelost met behulp van een blok LU decompositie. Deze methode kan aanzienlijk worden versneld door op te merken dat de matrix D diagonaliseerbaar is. De resulterende methode is nauw verwant met de techniek beschreven in [2].

[1] J.B.M. van Doremalen, " Two parallel queues with one-way overflow: a matrix structure approach ", In: Operations Research Proceedings 1984, Springer Verlag, Berlin, 1985.

[2] J.A. Morrison, " An overflow system in which queueing takes precedence ", Bell Systems Technical Journal, 60(1981)1-12.

VII

De analyse van ontwerpproblemen die geformuleerd worden met behulp van zogenaamde chance constraints, wordt veelal bemoeilijkt door het probleem van een herhaalde evaluatie van de constraints, zie bv. [1]. Het gebruik van benaderingsmethoden kan hier uitkomst bieden.

[1] J.B.M. van Doremalen, " The design of a system with two parallel queues and one-way overflow ", Memorandum COSOR 83-02, THE, Eindhoven, 1983.

VIII

Geautomatiseerde produktienetwerken en transportnetwerken vormen een uitdagend toepassingsgebied voor de theorie van netwerken van wachtrijen, zoals die in de afgelopen twintig jaar met name is ontwikkeld voor de performance evaluatie van communicatienetwerken en computersystemen.

IX

De invoering van het nieuwe uitlotingssysteem voor obligaties betekent met name voor kleine beleggers een variantie-reduktie in de opbrengst die niet door iedereen als prettig behoeft te worden ervaren.

X

Recente resulaten uit het onderzoek naar het sterfteproces van lichaamscellen tonen aan dat het gezegde " hardlopers zijn doodlopers " niet alleen figuurlijk dient te worden opgenomen.

[1] S. Hauser, " *Celslijtage en dementie* ", Intermediair, 22(1986)23-27.

XI

Naar analogie van de indeling in gewichtsklassen bij boksen en gewichtsheffen dient bij basketballen de indeling in lengteklassen overwogen te worden.

J.B.M. van Doremalen, 14 maart 1986.