Institut für Nachrichtenvermittlung und Datenverarbeitung

Universität Stuttgart

Prof. Dr.-Ing. A. Lotze

22. Bericht über verkehrstheoretische Arbeiten

Untersuchung von Systemen mit seriellem Warten

von

WOLFGANG KRÄMER

Institute of Switching and Data Technics

University of Stuttgart

Prof. Dr.-Ing. A. Lotze

22nd Report on Studies in Congestion Theory

Investigations of Systems with Queues in Series

by

WOLFGANG KRÄMER

1975

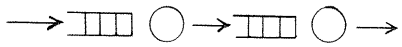
INVESTIGATIONS OF SYSTEMS WITH QUEUES IN SERIES

ABSTRACT

Predicting the performance of computer and switching systems has gained much importance, both for the planning of future and the modification of existing systems.

For this aim often queuing models are used which can be treated by means of probability theory, since the traffic within these systems may be described by stochastic processes.

Queues in series can be found by investigating the traffic flow in computers, switching and transmission systems, as well as in many other fields (production lines, supermarkets, etc.). They are characterized by a serial arrangement of service-units (○) in front of which the requests can wait for service on waiting places (□):



In this report two queues in series are analyzed with several distinct assumptions concerning system structures (number of service units and waiting places) and the traffic parameters (e.g. the distribution functions of the service times in both stages).

Preferably, such systems are investigated with one service unit in both stages, an infinite number of waiting places in the first stage and an intermediate buffer of infinite or finite length between the two service units. The arrival processes of

requests in the first stage are assumed to be Poisson processes, the distribution functions of the service times in both stages are mainly characterized by their first two moments.

New results have been derived concerning throughput, waiting times, etc., being useful for systems engineers to meet specific dimensioning performance requirements.

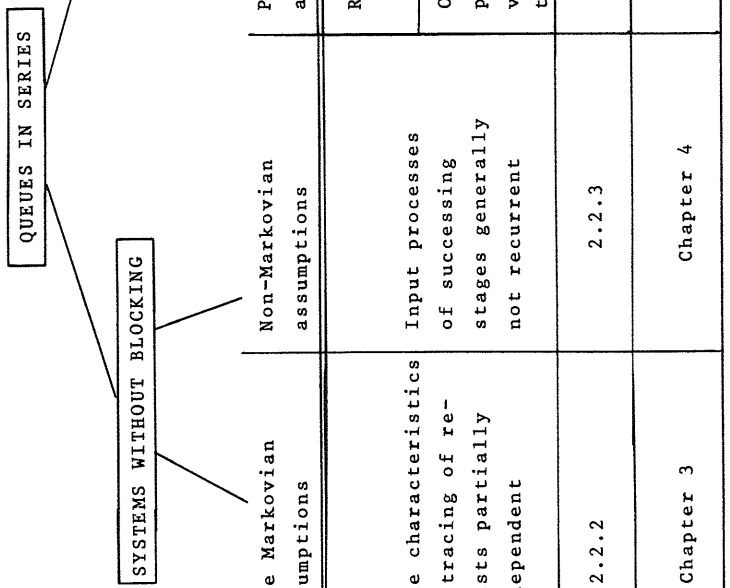
All results of this report concerning approximations are checked by extensive simulation results.

Fig. A presents a classification scheme of queues in series and may serve as a short orientation for this report.

In the following, a review about the various chapters of this report is given.

CHAPTER 1 INTRODUCTION TO SINGLE AND MULTI-STAGE QUEUING SYSTEMS

Beginning the introduction, the task of queuing theory is described. Using the example of single-stage queuing systems, important assumptions and prerequisites for a queuing analysis of such systems are cited. Then, the structure and the operating strategies of queues in series are explained. It is shown, how their traffic behaviour at various queuing disciplines can be described by characteristic traffic values (systems occupations, waiting times, etc.). This chapter is closed by some typical applications resulting from computers and from switching and transmission systems.



features				
	Pure Markovian assumptions	Non-Markovian assumptions	Pure Markovian assumptions	Non-Markovian assumptions
	Fate characteristics by tracing of requests partially independent	Input processes of successing stages generally not recurrent	Reduction of maximum throughput rate by blocking	
			Calculation always possible on principle via state probabilities	Only few solutions known
Survey in	2.2.2	2.2.3	2.3.1	2.3.2
New results in	Chapter 3	Chapter 4	-	Chapters 5, 6

Fig. A: Classification of queues in series and chapter hints

CHAPTER 2

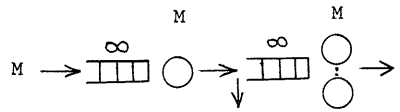
Chapter 2 gives an extensive survey about queues in series treated in the literature. This serves as a special introduction to the new investigations in Chapters 3-6 of this report and as a motivation for the systems treated there as well as for the methods of analysis used.

The most important analytical methods are presented together with their results.

This survey is systems-orientated and contains about 50 publications about this special topic.

CHAPTER 3 TRACING REQUESTS THROUGH THE WHOLE SYSTEM

In this chapter, two infinite queues in series under pure Markovian assumptions are investigated. In addition, probabilistic branching behind the first stage into several directions is admitted:



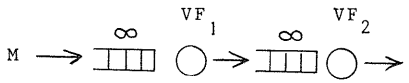
The first stage is a single-server stage with Poisson arrivals, the second is allowed to be a multi-server stage, both with negative exponentially distributed service times. Some of the results are also valid for arbitrary service time distribution functions in stage 2.

By tracing of special test-requests during their flow through the whole system, the fate of a request in stage 2 is determined as a function of all possible partial fates in stage 1 (Chapter 3.2). It is shown that the fate of a request in stage 2 is independent of the specific value (as long as it is >0) of its waiting time (not including service) in stage 1; that is to say, only dependent on the fact whether this request has waited in stage 1 or not. Therefore, in 3.3 the distribution function of the total waiting times of requests in this 2-stage system can be determined by convolution of conditional waiting time distribution functions.

Finally, in Chapter 3.4 also requests are investigated which have met a certain known number of requests (queue length) upon their arrival in stage 1. By considering all possible paths in a RANDOM WALK diagram, the number of requests met in stage 2 upon arrival there is determined, as well as the resulting waiting and flow times.

CHAPTER 4 CALCULATION OF 2 QUEUES IN SERIES WITHOUT BLOCKING WHEN INTERMEDIATE TRAFFICS ARE NON-RECURRENT

In this chapter two infinite queues in series with single servers are considered having Poisson arrivals in the first stage and arbitrary but hypoexponential distribution functions VF_1 and VF_2 of service times:

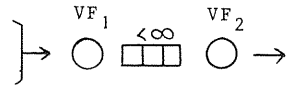


The first stage can be calculated exactly by the so-called Pollaczek-Khintchine formula. To calculate the mean total flow time, however, it is necessary to determine the mean waiting time in stage 2. Hereto it must be taken into account that the output process of the first unblocked stage is, in this general case, no longer a simple renewal process but one with dependencies between subsequent interarrival times. For arbitrary hypoexponentially distributed service times in both stages, the expected waiting time of all requests in the second stage is obtained by interpolation between known exact and new approximative results for all possible combinations VF_1-VF_2 with constant or negative-exponentially distributed service times. As a first step for this, the system with D-M is solved separately.

Known results are shown on page 107, the approximation formula can be found on page 117.

CHAPTER 5 MAXIMUM THROUGHPUT RATE FOR TWO SERVERS IN SERIES WITH FINITE INTERMEDIATE BUFFER

The system treated in Chapter 5 consists of two service units arranged in series and an intermediate buffer of finite length.



A very important traffic characteristic of such a system is the maximum throughput depending on the buffer size and the two service time distribution functions. To this, a simple approximation formula is derived which is applicable for arbitrary distribution functions of service times, only needing the first two moments (mean and variance).

The approximation is prepared in two steps:

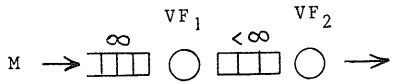
- first it is shown that the maximum throughput rate is exactly independent of the sequence of the two servers (this is done by setting up an equivalent closed cyclic queuing system);

- then it is shown that the maximum throughput can be assumed with very good accuracy to be independent of the exchange of the two types of service time distribution functions.

Known results are shown on page 122, the maximum throughput rate can be calculated according to formula (5.17), page 134.

CHAPTER 6 APPLICATION OF A CUTTING PRINCIPLE TO A SYSTEM OF TWO SINGLE SERVER STAGES WITH BLOCKING

In this chapter an approximation procedure is developed for the calculation of the non-saturated system appropriate to the system considered in Chapter 5, having Poisson arrivals in the first stage and arbitrary but hypoexponential service time distribution functions VF_1 and VF_2:



The calculation principle consists of the fact of splitting the two-stage system (despite the possibility of blocking) into 2 quasi independent single stages. The feature for the equivalent second stage is a finite waiting storage, the resulting loss at the same throughput rate is taken as a measure for the probability of blocking in the original system. Blocking is taken into consideration by enlarged occupation times of the first server.

The most important result is the mean total delay of requests, but also statements are made concerning the traffic characteristics of the single stages.

Related systems treated in the literature are shown on page 140, extensive simulation checks begin on page 156.

APPENDIX

In Appendix A1 all distribution functions investigated are listed up together with formulae for all moments and the so-called variation coefficients.

Appendix A2 is a summary of investigations concerning the influences of the moments of third and higher orders, and is a supplement to the results in Chapters 4, 5 and 6. It is shown explicitly, that the assumed 2-moments-approximation is useful in a wide range of application purposes.

INHALTSVERZEICHNIS Seite

LITERATURVERZEICHNIS

Allgemeines

|1| BHAT,U.N. Sixty years of queueing theory
Management Science 15(1969)6,B280-B294

|2| BHAT,U.N. Elements of applied stochastic processes
John Wiley,New York/London (1972)

|3| BROCKMEYER,E. HALSTRØM,H.L. JENSEN,A. The life and works of A.K.ERLANG
Acta Polytech.Scandinavica
Copenhagen (1960)

|4| COHEN,J.W. The single server queue
North-Holland,Amsterdam/London (1969)

|5| COHEN,J.W. Some aspects of queueing theory
Proc. 7th International Teletraffic Congress (ITC),Stockholm (1973),321/1-7

|6| COOPER,R.B. Introduction to queueing theory
The Macmillan Company,New York(1972)

|7| COX,D.R. SMITH,W.L. Queues
John Wiley,New York/London (1961)

|8| DOIG,A. A bibliography on the theory of queues
Biometrika 44(1957),490-514

|9| FELLER,W. Introduction to probability theory and its applications, Vol I,II
John Wiley,New York/London (1966)

|10| FERSCHL,F. Zufallsabhängige Wirtschaftsprozesse
Grundlagen und Anwendungen der Theorie der Wartesysteme
Physika-Verlag,Wien/Würzburg (1964)

|11| HERZOG,U. Verkehrsfluß in Datennetzen
Habilitationsschrift Univ. Stuttgart (1973)

|12| KLEINROCK,L. Communication nets
Mc Graw-Hill,New York/London (1964)

|13| KÜHN,P. Über die Berechnung der Wartezeiten in Vermittlungs- und Rechnersystemen
Dissertation Univ. Stuttgart (1972)

|14| LOTZE,A. Einführung in die Nachrichtenverkehrstheorie
Vorlesung an der Univ. Stuttgart

|15| RÉNYI,A. Wahrscheinlichkeitsrechnung
DVW,Berlin (1966)

|16| SAATY,T.L. Elements of queueing theory
Mc Graw-Hill,New York/London (1961)

|17| SAATY,T.L. Seven more years of queues
Naval Res.Log.Quart. 13(1966),447-476

|18| STÖRMER,H.et al Verkehrstheorie
R.Oldenbourg,München (1966)

|19| SYSKI,R. Introduction to congestion theory in telephone systems
Oliver and Boyd,Edinburgh/London (1960)

|20| TAKÁCS,L. Introduction to the theory of queues
Oxford University Press,New York (1962)

|21| ZIMMERMANN,G.O. STÖRMER,H. Wartezeiten in Nachrichtenvermittlungen mit Speichern
R.Oldenbourg,München (1961)

Simulation,Messung

|22| WAGNER,H. DIETRICH,G. Bestimmung der Verkehrsleistung von Wartesystemen durch künstlichen Fernsprechverkehr
Nachrichtentechn.Z. 17(1964),273-279

|23| HUBER,M. WAGNER,W. Simulation von Nachrichtenvermittlungssystemen. In:Nicht-numerische Informationsverarbtg.(Hrsgb.R.Gunzenhäuser)
Springer,Wien/New York (1968)

|24| IVERSEN,V.B. Analyses of real teletraffic processes based on computerized measurements
Ericsson Technics 29(1973)1,3-64

|25| KAMPE,G. KÜHN,P. LANGENBACH-BELZ, M. Simulation in der Nachrichtenverkehrstheorie, Problemstellungen und Programmiersprachen
GI-Workshop über rechnergestützte Simulation
KFK-Bericht 1845,Gesellschaft f.Kernforschg. Karlsruhe (1973)

|26| KÜMMERLE,K. Ein Vorschlag zur Berechnung der Vertrauensintervalle bei Verkehrstests
AEÜ 23(1969)10,507-511

|27| LOTZE,A. Über die statistische Sicherheit von Verkehrsmessungen
Nachrichtentechn.Z. 11(1958)1,5-7

Spezielle einstufige Systeme

|28| BOES,D.C. Note on the output of a queueing system
J.Appl.Prob. 6(1969),459-461

|29| BRANDT,G.J. Das preemptive Warteverlustsystem
Dissertation Univ.Stuttgart (1971)

|30| BURKE,P.J. The output of a queueing system
Op.Res. 4(1956),699-704

|31| CHANG,W. Output distribution of a single-channel queue
Op.Res. 11(1963),620-623

|32| COX,D.R. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables
Proc.Camb.Philos.Soc. 51(1955),433-441

|33| DALEY,D.J. The correlation structure of the output process of some single server queueing systems
Ann.Math.Stat. 39(1968),1007-1019

|34| FINCH,P.D. The effect of the size of the waiting room on a simple queue
J.Royal Stat.Soc.Ser B 20(1958)1,182-186

|35| FINCH,P.D. The output process of the queueing system M|G|1
J.Royal Stat.Soc.Ser B 21(1959)2,375-380

|36| JENKINS,J.H. On the correlation structure of the departure process of the $M|E_\lambda|1$ queue
J.Royal Stat.Soc.Ser B 28(1966),336-344

|37| KEILSON,J. The ergodic queue length distribution for queueing systems with finite capacity
J.Royal Stat.Soc.Ser B 28(1966),190-201

|38| KENDALL,D.G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain
Ann.Math.Stat. 24(1953),338-354

|39| KRÖNER,G. Approximative Berechnung der mittleren Wartezeiten im Wartesystem GI|G|1
Studienarbeit Nr.425 am Institut für Nachrichtenvermittlung und Datenverarbeitung Univ.Stuttgart (1974)

|40| LINDLEY,D.V. The theory of queues with a single server
Proc.Camb.Philos.Soc. 48(1952),277-289

|41| LITTLE,J.D.C. A proof for the queueing formula: $L=\lambda \cdot W$
Op.Res. 9(1961),383-387

|42| LOTZE,A. Berechnung der Verkehrsgrößen im Wartezeitsystem aus den Verkehrsgrößen eines Verlustsystems
Fernmeldetechn.Zeitschr. 7(1954),443-453

|43| MIRASOL,N.M. The output of an $M|G|\infty$ queueing system is Poisson
Op.Res. 11(1963),282-284

|44| SMITH,W.L. On the distribution of queueing times
Proc.Camb.Philos.Soc. 49(1953),449-461

|45| WAGNER,W. Über ein kombiniertes Warte-Verlustsystem mit Prioritäten
Dissertation Univ.Stuttgart (1968)

Warteschlangen in Rechnern

|46| BUZEN,J.P. Queueing network models of multiprogramming
Thesis Harvard Univ. Cambridge,Mass(1971)

|47| CHANG,W. Single server queueing processes in computing systems
IBM Systems Journal 9(1970)1,36-71

|48| COFFMAN,E.G.
DENNING,P.J. Operating systems theory
Prentice Hall,Englewood Cliffs,N.J.(1973)

|49| HERZOG,U.
KÜHN,P.
ZEH,A. Klassifizierung und Analyse von Verkehrsmodellen für das Ablaufgeschehen in Rechnersystemen
Nachrichtentechn.Fachber. 44(1972),181-198

|50| HERZOG,U.
KRÄMER,W.
KÜHN,P.
WIZGALL,M. Analyse von Betriebssystem-Modellen für Rechnersysteme mit Multiprogramming und Paging
NTG-Fachtagung "Struktur und Betrieb von Rechensystemen" Braunschweig
Lecture Notes in Computer Science 8
Springer,Berlin/New York (1974),266-288

|51| KÜHN,P. Parallel waiting queues in real-time computer systems
Nachrichtentechn.Z. 23(1970) 11,576-582

|52| KÜSPERT,H.J.
MARTE,G. Optimale Rechenzeit-Zuteilung bei einem Teilnehmer-Rechensystem mit jeweis einer Aufgabe im Arbeitsspeicher
Elektron.Rechenanl. 12(1970)3,155-162

|53| LANGENBACH-BELZ, M. Getaktete Wartesysteme bei Rechnern und zentralgesteuerten Nachrichtenvermittlungsanlagen
Dissertation Univ.Stuttgart (1973)

|54| REISER,M.
KOBAYASHI,H. The effects of service time distributions on system performance
Proc.IFIP 74,North-Holland (1974),230-234

|55| SWOBODA,J. Verkehrsfragen innerhalb einer Rechenanlage
Elektron.Rechenanl. 12(1970)5,249-252

Mehrstufige Wartesysteme

|56| AVI-ITZHAK,B.
YADIN,M. A sequence of two servers with no intermediate queue
Management Science Ser A 11(1965),553-564

|57| AVI-ITZHAK,B. A sequence of service stations with arbitrary input and regular service times
Management Science Ser A 11(1965),565-571

|58| O'BRIEN,G.G. The solution of some queueing problems
J.Soc.Ind.Appl.Math. 2(1954),133-142

|59| BURKE,P.J. The dependence of delays in tandem queues
Ann.Math.Stat. 35(1964),874-875

|60| BURKE,P.J. The output-process of a stationary M|M|s queueing system
Ann.Math.Stat. 39(1968),1144-1152

|61| BURKE,P.J. The dependence of sojourn times in tandem M|M|s queues
Op.Res. 17(1969),754-755

|62| BURKE,P.J. Output processes and tandem queues
Proc.Symp.on Computer-Communications Networks and Teletraffic,Brooklyn
Polytechnic Press (1972),419-428

|63| CHANG,W. Two servers in series with Erlang input
Proc.Symp.on Computer-Communications Networks and Teletraffic,Brooklyn
Polytechnic Press (1972),409-418

|64| COHEN,J.W. On the queueing process of lanes Philips Techn.Report (1956)

|65| FRIEDMAN,H. Reductions methods for tandem queueing systems Op.Res. 13 (1965),121-131

|66| GHOSAL,A. Queues in series J.Royal Stat.Soc.Ser B 24(1962),359-363

|67| HATCHER,J.M. The effect of internal storage on the production rate of a series of stages having exponential service times AIIE Transactions 1(1969),150-156

|68| HAYDON,B.J. System states for a series of finite queues Op.Res. 20(1972)5,1137-1141

|69| HILDEBRAND,D.K. Stability of finite queue,tandem server systems J.Appl.Prob. 4(1967),571-583

|70| HILDEBRAND,D.K. On the capacity of tandem server,finite queue service systems Op.Res. 16(1968),72-82

|71| HILLIER,F.S. BOLING,R.W. The effect of some design factors on the efficiency of production lines with variable operation times J.Indust.Eng. 17(1966),651-658

|72| HILLIER,F.S. BOLING,R.W. Finite queues in series with exponential or Erlang service times.A numerical approach. Op.Res. 15(1967),286-303

|73| HUNT,G.C. Sequential arrays of waiting lines Op.Res. 4(1956),674-683

|74| JACKSON,J.R. Networks of waiting lines Op.Res. 5(1957),518-521

|75| JACKSON,R.R.P. Queueing systems with phase-type service Oper.Res.Quart. 5(1954),109-120

|76| JACKSON,R.R.P. Random queueing processes with phase-type service J.Royal Stat.Soc.Ser B 18(1956),129-132

|77| KRÄMER,W. TALPALARU,R. Näherungsweise Berechnung eines 2-stufigen Wartesystems mit Übergabezeiten und einer Bedienungseinheit je Stufe Monographie des Instituts NVDV Univ.Stuttgart (1973)

|78| KRÄMER,W. SCHUON,G. Berechnung von 2-stufigen Wartesystemen mit Übergabezeiten und mehreren Bedienungseinheiten in der 2.Stufe. Monographie des Instituts NVDV Univ.Stuttgart (1973)

|79| LEE,L. Waiting time distributions for networks of delay systems Proc. 5th International Teletraffic Congress (ITC),New York (1967),310-317

|80| LOYNES,R.M. The stability of a system of queues in series Proc.Camb.Philos.Soc. 60(1964),569-574

|81| LOYNES,R.M. On the waiting time distribution for queues in series J.Royal Stat.Soc. Ser B 27(1965),491-496

|82| MAKINO,T. On the mean passage time concerning some queueing problems of the tandem type J.Opns.Res.Japan 7(1964),17-47

|83| MAKINO,T. On a study of output distribution J.Opns.Res.Japan 8(1966),109-133

|84| MASTERSON,G.E. SHERMAN,S. On queues in tandem Ann.Math.Stat. 31(1960),239

|85| MASTERSON,G.E. SHERMAN,S. On queues in tandem Ann.Math.Stat. 34(1963),300-307

|86| MORSE,P.M. Queues,inventories and maintenance John Wiley,New York/London (1958)

|87| MUTH,E.J. The production rate of a series of work stations with variable service times Florida Univ.,Gainesville,Dept.of industrial and systems engineering Contract DAHC04-68-C-0002 (1971)

|88| NELSON,R.T. Waiting-time distributions for application to a series of service centers Op.Res. 6(1958),856-862

|89| NEUTS,M.F. Two queues in series with a finite intermediate waiting room J.Appl.Prob. 5(1968),123-142

|90| NEUTS,M.F. Two server in series,studied in terms of a Markov renewal branching process Adv.Appl.Prob. 2(1970),110-149

|91| PACK,C.D. The effects of multiplexing on a computer--communications system Comm.ACM 16(1973)3,161-168

|92| PATTERSON,R.L. Markov processes occurring in the theory of traffic flow through an N-stage stochastic service system J.Indust.Eng. 15(1964)4,188-193

|93| PRABHU,N.U. Transient behaviour of a tandem queue Management Science 13(1967),631-639

|94| REICH,E. Waiting times when queues are in tandem Ann.Math.Stat. 28(1957),768-772

|95| REICH,E. Note on queues in tandem Ann.Math.Stat. 34(1963),338-341

|96| REICH,E. Departure processes Proc.Symp.on Congestion Theory The Univ.of North Carolina Press,Chapel Hill, August (1964)

|97| RÖCK,H. Exakte Berechnung von 2-stufigen Bedienungssystemen mit seriellem Warten bei Markoff'schen Voraussetzungen und beschränkter Wartemöglichkeit Studienarbeit Nr.358 am Institut für Nachrichtenvermittlung und Datenverarbeitung Univ.Stuttgart (1972)

|98| SAATY,T.L. Stochastic network flows: Advances in networks of queues Proc.Symp.on Congestion Theory The Univ.of North Carolina Press,Chapel Hill, August (1964)

|99| SACKS,J. Ergodicity of queues in series Ann.Math.Stat. 31(1960),579-588

|100| STANGE,K. Zwei in Reihe angeordnete Schalter (mit einer Warteschlange dazwischen) bei Exponentialverteilung der Ankünfte und Abgänge ZAMM 42(1962),T81-T83

|101| STANGE,K. Zwei in Reihe angeordnete Schalter (mit einer Warteschlange dazwischen) bei Exponentialverteilung der Ankünfte und Abgänge Unternehmensforschung 6(1962),101-124

|102| SUZUKI,T. Two queues in series J.Opns.Res.Japan 5(1963)4,149-155

|103| SUZUKI.T. On a tandem queue with blocking J.Opns.Res.Japan 6(1964)3,137-157

|104| SUZUKI,T. Ergodicity of a tandem queue with blocking J.Opns.Res.Japan 7(1964)2,68-75

|105| SWOBODA,J. ROSENBOHM,W. Modell für den Befehlsablauf in einer Rechenanlage: Eine Serverkette mit vorgebbarer Varianz der Belegungsdauern GI-Jahrestagung Hamburg Lecture Notes in Computer Science 1 Springer,Berlin/New York (1973),314-326

|106| WDOWIN,A.A. ИССЛЕДОВАНИЕ НЕКОТОРЫХ ДВУХФАЗНЫХ СИСТЕМ МАССОГО ОБСЛУЖИВАНИЯ (Untersuchung einiger Systeme mit 2 Phasen) Arbeiten d.zentr. Forschungsinst. f. Automatisierung 12(1965),237-257

Sonstige Wartesysteme

|107| GORDON,W.J. NEWELL,G.F. Closed queuing systems with exponential servers Op.Res. 15(1967),254-265

|108| MILCH,P.R. WAGGONER,M.H. A random walk approach to a shutdown queueing system SIAM J.Appl.Math. 19(1970),103-115

BEZEICHNUNGEN

Alle Bezeichnungen werden mindestens beim erstmaligem Auftreten erklärt.
Die Punkte 1-6 enthalten Bezeichnungen die sich auf alle Kapitel beziehen,in 7-10 sind jeweils kapitelspezifische Abkürzungen aufgeführt.

1 Allgemeine Größen

t	Zeit
t_0, t_1	feste Zeitdauern
P(..)	Wahrscheinlichkeit für das in der Klammer bezeichnete Ereignis
E(..)	Erwartungswert für die in der Klammer bezeichnete Zufallsvariable
A|B	A unter der Bedingung B
Var(..)	Varianz einer Zufallsvariablen
Cov(..,..)	Kovarianz zweier Zufallsvariablen
r(..,..)	Korrelationskoeffizient zweier Zufallsvariablen
m_j	gewöhnliches Moment j.Ordnung
σ^2	$=m_2-m_1^2$ Varianz } mit Index { A für Ankunftsprozeß; H für Bedienungsprozeß; D für Ausgangsprozeß; W für Wartezeiten; B für Blockierzeiten
c	$=\sigma/m_1$ Varianzkoeffizient } mit Index { A für Ankunftsprozeß; H für Bedienungsprozeß; D für Ausgangsprozeß; W für Wartezeiten; B für Blockierzeiten
f(..)	Funktion von ..
*	Kennzeichnung von Ersatzgrößen
..Y	Index zur Kennzeichnung von Größen erfolgreicher Anforderungen

2 Strukturparameter

n	Zahl der Bedienungseinheiten (BE)
n_i	Zahl der BE in Stufe i
s	Zahl der Warteplätze
s_i	Zahl der Warteplätze im Wartespeicher der Stufe i
m	Zahl der Stufen eines mehrstufigen Wartesystems

Bei den meisten nachfolgenden Bezeichnungen gibt bei Anwendung auf mehrstufige Systeme ein eventueller zusätzlicher Index die Stufe an,auf die sich die Größe bezieht.Größen ohne diesen Index beziehen sich bei mehrstufigen Systemen auf das Gesamtsystem.

3 Zufallsvariable

T_A	Ankunftsabstand (interarrival time)
T_H	Bedienungszeit (service time), Belegungsdauer (holding time)
T_W	Wartezeit (waiting time)
T_B	Blockierzeit (blocking time)
T_F	Durchlaufzeit (flow time)
T_{WB}	Verzögerungszeit (Summe aus Warte- und Blockierzeit)
T_D	Abgangsabstand (interdeparture time)
X	Zahl der Anforderungen im System (BE u. Wartespeicher)
X_A	Zahl der angetroffenen Anforderungen bei Ankunft (arrival) einer Anforderung
X_D	Zahl der zurückgelassenen Anforderungen bei Abgang (departure) einer Anforderung

4 Verkehrsparameter

λ	Ankunftsrate
λ_Y	Rate der erfolgreichen Anforderungen
λ_{max}	$=\lambda_{Y_{max}}$ maximal verarbeitbare Rate (Durchsatz)
h	$=E(T_H)$ mittlere Bedienungszeit
ε	Enderate einer negativ-exponentiell verteilten Bedienungsphase ($\varepsilon=1/h$ bei neg.-exp. Bedienungszeiten)
μ	$= n/h$ Enderate einer Gruppe von BE (Bündel) ($\mu = n\varepsilon$ bei negativ-exponentiellen Bedienungszeiten)
A	$=\lambda \cdot h$ Angebot (Erl)
ϑ	$= A/n$ Angebot pro BE, Ausnutzung
B	Verlustwahrscheinlichkeit
W	Wartewahrscheinlichkeit
$E_{2n}(A)$	2.Erlang sche Formel (W im System M|M|n)
p_B	Blockierwahrscheinlichkeit
p_{WB}	Verzögerungswahrscheinlichkeit
Y	$=A(1-B)$ Verkehrswert oder Belastung (Erl)
Y_B	Blockierbelastung
Ω	Mittlere Warteschlangenlänge (Wartebelastung)
t_W	Mittlere Wartezeit der Wartenden $=E(T_W \mid T_W>0)$
t_B	Mittlere Blockierzeit der Blockierten $=E(T_B \mid T_B>0)$
t_{WB}	Mittlere Verzögerungszeit der Verzögerten $=E(T_{WB} \mid T_{WB}>0)$

p(x)	$=P(X=x)$ (absolute) stationäre Zustandswahrschkt.
$p(x_1,..x_m)$	stationäre Wahrscheinlichkeit für Zustandsmuster $\{x_1,..x_m\}$ in einem m-stufigen Wartesystem

5.Verteilungsfunktionen (VF)

$A(\leqslant t)$	$=P(T_A \leqslant t)$ VF der Ankunftsabstände
$H(\leqslant t)$	$=P(T_H \leqslant t)$ VF der Bedienungszeiten
$W(\leqslant t)$	$=P(T_W \leqslant t)$ VF der Wartezeiten
$B(\leqslant t)$	$=P(T_B \leqslant t)$ VF der Blockierzeiten
$F(\leqslant t)$	$=P(T_F \leqslant t)$ VF der Durchlaufzeiten
$W(>t)$	$=P(T_W>t)$ (komplementäre) VF der Wartezeiten bezogen auf alle Anforderungen, $W(>0)=W$
$W_W(>t)$	$=P(T_W>t \mid T_W>0)$ (komplementäre) VF der Wartezeiten bezogen auf die Wartenden, $W_W(>0)=1$
$f_A(t)$	$=A'(\leqslant t)$ Dichtefunktion der Ankunftsabstände
$f_H(t)$	$=H'(\leqslant t)$ Dichtefunktion der Bedienungszeiten
$f_W(t)$	$=W'(\leqslant t)$ Dichtefunktion der Wartezeiten
k	Zahl der Phasen oder Zweige einer VF
p_i	Verzweigungswahrscheinlichkeit bei VF mit k Zweigen
h_i	Mittlere (Teil-)Bedienungszeit in Zweig oder Phase i einer Bedienung
$p(i,t_o)$	(diskrete) Poisson-VF (Wahrscheinlichkeit für i Poisson-Ereignisse während t_o)

Kennzeichnung des Typs einer VF:

VF_i	VF der Bedienungszeiten in Stufe i
G	Beliebige VF (general)
GI	Beliebige VF unabhängiger Ankunftsabstände (general independent input)
M	Negativ-exponentielle VF (Markovian)
D	Konstante VF (deterministic)
E_k	Erlang-k-VF
H_k	Hyperexponentielle VF k. Ordnung
P_k	Mehrpunkt-VF mit k Punkten
DM	"verschobene negativ-exponentielle VF"

6 Kennzeichnung von Systemen (analog KENDALL |38|)

6.1 Einstufige Systeme

Ankunftsprozess|Bedienungsprozess|Zahl der BE

z.B. $M|M|n$, $D|E_2|1$, $GI|G|1$

Bei endlich großem Speicher wird die Zahl s der Warteplätze angefügt, z.B. $M|G|1\text{-}s$, $D|E_k|n\text{-}s$

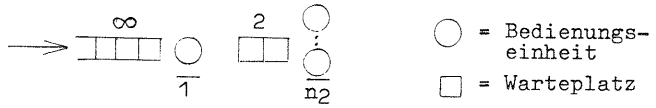
6.2 Zweistufige Systeme

$$\text{Ankunftsprozess} \xrightarrow{s_1} \underbrace{VF_1|n_1}_{\text{Stufe 1}} \xrightarrow{s_2} \underbrace{VF_2|n_2}_{\text{Stufe 2}}$$

z.B.

$$M \xrightarrow{\infty} D|1 \xrightarrow{2} M|n_2$$

zugehörige Systemstruktur:



○ = Bedienungseinheit

□ = Warteplatz

Für eine gegebene Systemstruktur (2-stufig) bezeichnet VF_1-VF_2 die beiden VF-Typen der Bedienungszeiten in Stufe 1 und 2, z.B. D-M in obigem System.

Bei mehrstufigen Wartesystemen ist bei maximaler Durchsatzrate bei endlich großen Zwischenspeichern die 1.Stufe durch Bedienung und Blockierung immer voll belegt (Zugangsprozeß als "Holprozeß" aus unendlich großem Reservoir):

$$\rightarrow D|1 \xrightarrow{2} M|n_2 \qquad (\lambda_Y = \lambda_{max})$$

Obige Kennzeichnungen werden hier auch sinngemäß auf beliebigstufige Systeme angewandt.

7 Spezielle Größen zur Schicksalsberechnung in Kapitel 3

η Wahrscheinlichkeit, daß eine Anforderung in die betrachtete Richtung geht

c Präfix zur Kennzeichnung von Test-Anforderungen
- c=0: Anforderung hat in Stufe 1 nicht gewartet
- c=1: Anforderung mit bekannter/unbekannter Wartezeit $T_{W1}>0$

$p_W(j,t_o)$ Wahrscheinlichkeit, daß eine Anforderung mit Wartezeit $T_W=t_o$ bei Ankunft j Anforderungen antraf

${}_cp_2^s(x|t_1)$ bedingte Zustandswahrscheinlichkeit in Stufe 2 kurz nach Beginn der Bedienungszeit $T_{H1}=t_1$ einer durch c gekennzeichneten Test-Anforderung in Stufe 1 (Startwahrscheinlichkeit)

${}_cp_2^a(x|t_1)$ bedingte Zustandswahrscheinlichkeit in Stufe 2 kurz vor Ende der Bedienungszeit $T_{H1}=t_1$ einer durch c gekennzeichneten Test-Anforderung (Antreffwahrscheinlichkeit)

${}_cF$ Faktor für bedingte Wartewahrscheinlichkeiten in Stufe 2

$F(\eta,s_2)$ Relationsfaktor für bedingte Wartewahrscheinlichkeiten in Stufe 2

Größen des RANDOM WALK:

p_i Wahrscheinlichkeit für nächstes Belegungsende in Stufe i; i=1,2

$\boxed{i_1,i_2}$ Durchlaufszustand für Test-Anforderung. enthält nur alle relevanten Anforderungen von $\{x_1,x_2\}$

$p_F(i_1,i_2)$ Wahrscheinlichkeit für Durchlaufszustand $\boxed{i_1,i_2}$

d_i "Distanzen" eines Weges i=0,1,2 =Zahl der Übergänge mit Wahrschkt. p_i (p_o=1)

$p_2(x|x_1,x_2)$ Wahrscheinlichkeit, daß eine Test-Anforderung, die bei Ankunft in Stufe 1 den Systemzustand $\{x_1,x_2\}$ vorfand, x Anforderungen bei Ankunft in Stufe 2 dort antrifft

$p_2(x|x_1)$ Wahrscheinlichkeit, daß eine Test-Anforderung, die bei Ankunft in Stufe 1 x_1 Anforderungen vorfand, x Anforderungen bei Ankunft in Stufe 2 antrifft

8 Spezielle Größen zu Kapitel 4

$E(T_W)_{GI|G|1}$ Erwartungswert der Wartezeit aller Anforderungen im System GI|G|1

$E(T_{W2})_{VF1\text{-}VF2}$ Erwartungswert für die Wartezeit in Stufe 2 eines 2-stufigen Systems mit VF_1-VF_2

o Index zur Kennzeichnung e.Größe für $\lambda\to 0$

$F_{oVF1\text{-}VF2}$ Grenzverhältnis von mittl. Wartezeiten für $\lambda\to 0$

9 Spezielle Größen zu Kapitel 5

$H_{max}(\leq t)$ VF der Zufallsvariable $\max(T_{H1}, T_{H2})$

$f_{Hmax}(t)$ zugehörige Dichtefunktion

$\lambda_{maxVF1-VF2}$ maximaler Durchsatz für angegebene VF-Kombination

$\lambda_{max}(c^2_{H1}+c^2_{H2})$ maximaler Durchsatz für angegebene Summe der Quadrate der Varianzkoeffizienten

$t_{BVF1-VF2}$ mittlere Blockierzeit der Blockierten für VF_1-VF_2

10 Spezielle Größen zu Kapitel 6

$F_{VF1-VF2}$ Adaptionsfaktoren für Blockierwahrscheinlichkeit

P_x (stationäre) Zustandswahrscheinlichkeit an den Regenerationspunkten

q_i Wahrscheinlichkeit, daß während einer zufälligen Bedienungszeit i Anforderungen in einem Poisson-Prozess mit Rate λ ankommen

1 EINFÜHRUNG IN EIN- UND MEHRSTUFIGE WARTESYSTEME

In Datenverarbeitungsanlagen, Nachrichtenvermittlungssystemen, Fertigungsstraßen, wie auch Systemen des täglichen Lebens tritt serielles Warten auf. Dabei benötigt eine eintreffende Anforderung, auch "Kunde" oder "Ruf" genannt, zu ihrer vollständigen Bearbeitung mehrere zeitlich nacheinander ablaufende Bearbeitungs- oder Bedienungsphasen in verschiedenen seriell angeordneten, unabhängig arbeitenden Bedienungseinheiten (BE). Einfache Beispiele hierfür sind Programme in Datenverarbeitungsanlagen, die zunächst auf die Eingabe, dann auf einen zentralen Prozessor warten und schließlich eine Ausgabeeinheit benötigen. Auch können die BE als Übertragungsleitungen in Datennetzen mit sog. "Message- oder Packet-Switching" aufgefaßt werden, wodurch auch einfache Teilkonfigurationen eines Netzes dargestellt werden können.

Im 1. Kapitel dieser Arbeit, die sich mit der exakten und approximativen Berechnung solcher Systeme befaßt, soll eine kurze Einführung in ein- und mehrstufige Wartesysteme gegeben werden. Dazu werden zunächst die Aufgaben der Bedienungstheorie erläutert. Anhand sog. einstufiger Wartesysteme werden wichtige Grundlagen und Voraussetzungen für eine verkehrstheoretische Behandlung solcher Systeme geschildert. Daran anschließend werden der Aufbau (Struktur) und die Funktionsweise (Betriebsstrategien) von mehrstufigen Wartesystemen erklärt. Es wird gezeigt, wie deren Verkehrsverhalten bei verschiedenen Betriebsstrategien durch charakteristische Verkehrsgrößen (Systembelastungen, Wartezeiten...) beschrieben werden kann.

Den Abschluß dieses Kapitels bilden einige typische Anwendungsbeispiele aus dem Bereich der Rechner- und Nachrichtenverkehrstheorie.

1.1 Allgemeines zur Bedienungstheorie

Bedienungssysteme sind Systeme verschiedenster Art, die irgendwie eintreffende und entstehende "Bedienungswünsche" oder "Anforderungen" befriedigen. Durch den im allgemeinen zufälligen (stochastischen) Charakter der Ankünfte und Bedienungszeiten kann es aus Mangel

an BE zu Verkehrshemmungen kommen,wenn momentan mehr Anforderungen an das System gestellt werden,als dieses maximal gleichzeitig befriedigen kann.Dabei können,je nach Art des Bedienungssystems, die Anforderungen in mehr oder weniger großer Zahl auf ihre Bedienung warten.
Ist diese sog. Warteschlange in ihrer Länge nicht begrenzt,so spricht man von einem reinen Wartesystem;kann überhaupt keine Anforderung warten,geht also eine eintreffende Anforderung verloren,wenn alle erreichbaren BE belegt sind,so liegt ein reines Verlustsystem vor. Steht nur ein endlich großer Wartespeicher zur Verfügung,so wird dieses System als kombiniertes Warte--Verlustsystem bezeichnet.
Beispiele für solche Systeme findet man an Schaltern und Kassen von Supermärkten,bei Fertigungsprozessen,im Straßenverkehr ,in Rechnersystemen auf Programm- und Befehlsebene,in Vermittlungssystemen und in vielen anderen Bereichen.
Bei der verkehrstheoretischen Behandlung solcher Bedienungssysteme werden in erster Linie analytische Berechnungsmethoden angewandt,die auf statistischen Gesetzmäßigkeiten im behandelten System aufbauen und so Aussagen über dessen statistisches Verhalten liefern |2-4,7,9,10,13-16,18-21|.
Häufig wird auch die Simulation auf einem Digitalrechner benutzt und das System mit seinem zufallsmäßigen Verkehr nachgebildet, wobei man gleiche Aussagen,jedoch mit einem gewissen Vertrauensintervall behaftet,gewinnen kann |22,23,25,26|.Diese Simulation ist notwendig,um bei Systemen,deren exakte Berechnung entweder nicht bekannt ist oder zu aufwendig wäre,überhaupt verkehrstheoretische Aussagen machen zu können oder auch um die Güte einer Näherungslösung zu bestimmen.
Darüber hinaus besteht die Möglichkeit,Aussagen über das Verhalten eines bestehenden Systems direkt durch Messung zu erhalten |24,27|.
Die Aufgabe der Bedienungs- bzw. Verkehrstheorie besteht nun darin,optimale Systemstrukturen und Betriebsstrategien zu entwickeln,so daß sich die Anforderungen gegenseitig möglichst wenig behindern und trotzdem die BE gut ausgelastet sind.Dabei gibt es je nach System verschiedene Zielsetzungen,die zu unterschiedlichen Lösungen führen können. Wird z.B. bei einem Rechnersystem auf maximalen Durchsatz (Zahl der bearbeiteten

Anforderungen pro Zeiteinheit) abgehoben,wie z.B.bei reinem Stapelbetrieb,so ergeben sich u.U. andere Systemstrukturen und insbesondere andere Betriebsstrategien zur Zuteilung der Betriebsmittel als z.B. im Falle eines Prozeßrechners,der eine schnelle Reaktionszeit erfordert.
Mit Hilfe der Verkehrstheorie werden nun allgemein Fragen der Analyse und Synthese solcher Systeme behandelt.Dabei stehen Fragen im Vordergrund nach der Zahl und Anordnung der BE (Systemstruktur),den Zuteilungsregeln für Warteplätze und BE (Betriebsstrategien),dem speziellen statistischen Charakter des Verkehrs (Verkehrsart) und der Zahl der pro Zeiteinheit ankommenden bzw. abgefertigten Anforderungen (Verkehrsintensität).
Bei der Analyse werden bei gegebener Systemstruktur,Betriebsstrategie und Verkehrsart

- die Verkehrsgüte für eine bestimmte Verkehrsintensität oder
- die zulässige Verkehrsintensität bei vorgeschriebener Verkehrsgüte

bestimmt.
Bei der Synthese werden bei gegebener Verkehrsart,Verkehrsintensität und Verkehrsgüte

- die Systemstruktur und die Betriebsstrategien

ermittelt.
Unter zusätzlicher Berücksichtigung von Kostenfunktionen kann eine Zielfunktion gebildet werden.Mit deren Hilfe ist die echte Optimierung eines solchen Systems möglich.Die Aussagen über die Verkehrsgüte können jedoch nur mit Hilfe verkehrstheoretischer Untersuchungen erfolgen,die also für die optimale Auslegung stochastischer Systeme eine wichtige Voraussetzung darstellen.
Bevor die in dieser Arbeit behandelten mehrstufigen Wartesysteme beschrieben werden,soll zunächst in 1.2 eine Einführung in sog. einstufige Systeme gegeben werden.Anhand dieser werden dann die obigen allgemeineren Ausführungen konkretisiert.

1.2 Einstufige Wartesysteme

Unter einstufigen Wartesystemen werden solche Systeme verstanden, bei denen eine Anforderung durch genau 1 BE fertig bedient werden kann, während bei mehrstufigen Anordnungen mehrere Bedienungsphasen durch verschiedene BE notwendig sind.
In diesem Abschnitt wird zur Vorbereitung auf mehrstufige Anordnungen eine Einführung in einstufige Wartesysteme gegeben. Dabei werden die verschiedenen Systemparameter beschrieben, die charakteristischen Verkehrsgrößen definiert und einige wichtige Berechnungsmethoden aufgeführt. Aus Umfangsgründen kann dies nur in kompakter Form geschehen, dafür werden in 1.2.4 Hinweise auf zugehörige grundlegende und weiterführende Literatur gegeben.

1.2.1 Beschreibungsweise und Voraussetzungen

Die notwendigen Angaben zur Beschreibung eines verkehrstheoretischen Modells lassen sich in folgende 3 wichtige Parameterarten gliedern:

- Strukturparameter — zur Kennzeichnung der Struktur des Systems
- Verkehrsparameter — zur Charakterisierung der Ankunfts- und Bedienungsprozesse
- Betriebsparameter — bezüglich der Strategien bei der Vergabe der Betriebsmittel.

● Die Strukturparameter eines Systems von Warteschlangen kennzeichnen im allg. die Zahl und Anordnung der Betriebsmittel, das sind Bedienungseinheiten (BE) und Warteplätze. Dies soll anhand des einfachen Beispiels eines einstufigen Wartesystems verdeutlicht werden, vgl. Bild 1.1 .

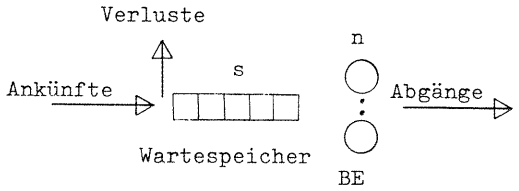


Bild 1.1 Einstufiges Wartesystem

Dieses besteht aus einer Gruppe von n BE und einem Wartespeicher mit s Warteplätzen. Die ankommenden Anforderungen belegen bei Ankunft eine freie BE oder warten, sofern noch mindestens ein Warteplatz frei ist auf ihre Abfertigung, bzw. "gehen verloren", falls kein Platz mehr im Wartespeicher ist. Je nach Größe des Wartespeichers handelt es sich bei

- $s \to \infty$ um ein reines Wartesystem
- $0 < s < \infty$ um ein kombiniertes Warte- Verlustsystem
- $s = 0$ um den besonders für Vermittlungssysteme wichtigen Fall des reinen Verlustsystems.

Als weiteres Strukturmerkmal kann die Zahl der Verkehrsquellen gelten, in denen die Anforderungen in zufälligen Abständen erzeugt werden. Ist diese Zahl nicht groß im Vergleich zur Zahl der BE, ist die Ankunftsrate spürbar von der momentanen Zahl von Anforderungen im System (Zahl der belegten Quellen) abhängig. Bei großer Zahl von Quellen hat man praktisch eine konstante Ankunftsrate der Anforderungen. Letzteres wird bei allen Strukturen dieser Arbeit angenommen.
Im obigen Beispiel wird vorausgesetzt, daß jede BE prinzipiell von jeder Anforderung belegt werden kann. In diesem Falle spricht man von "vollkommener Erreichbarkeit". Dagegen werden bei Vermittlungssystemen zur Ersparnis von Koppelpunkten häufig ein- oder mehrstufige Koppelanordnungen mit unvollkommener Erreichbarkeit ("Mischungen", evtl. "Linksysteme") verwendet. Dabei können die aus verschiedenen Richtungen oder Teilgruppen eintreffenden "Rufe" jeweils nur eine bestimmte Auswahl von "Leitungen" des "Bündels" absuchen. Im Rahmen dieser Arbeit sollen alle BE einer Gruppe erreichbar und gleichartig sein.

● Mit den Verkehrsparametern werden die statistischen Eigenschaften der Ankunfts- und Bedienungsprozesse (Verkehrsart und Verkehrsintensität) charakterisiert. Durch viele Messungen z.B. in Rechnern oder Vermittlungssystemen hat man bestätigen können, daß die Verkehre, die in sich gesehen determiniert ablaufen, in Bezug auf das System gesehen jedoch zufälligen Charakter besitzen und deshalb mit Mitteln der Wahrscheinlichkeitstheorie beschrieben werden können. Dazu werden zufällig schwankende Zeiten (kontinuierliche Zufallsvariable) durch Verteilungs-

funktionen (VF) gekennzeichnet.Diese VF geben an,mit welcher Wahrscheinlichkeit die betreffende Zufallsvariable höchstens gleich der Zeit t ist (z.B. die Ankunftsabstände zwischen zwei aufeinanderfolgenden Anforderungen oder die Bedienungszeiten von Anforderungen).

Bezeichnet T_A den zufälligen Ankunftsabstand zweier aufeinanderfolgender Anforderungen,so sei die zugehörige VF

$$A(\leq t) \stackrel{def}{=} P(T_A \leq t) \qquad (1.1)$$

Für eine zufällige Bedienungszeit T_H sei analog

$$H(\leq t) \stackrel{def}{=} P(T_H \leq t) \qquad (1.2)$$

Als wichtigste VF gilt die sog. negativ-exponentielle VF,sowohl für Ankunftsabstände

$$A(\leq t) = 1 - e^{-\lambda t}$$

als auch für Bedienungszeiten

$$H(\leq t) = 1 - e^{-\varepsilon t} \qquad (1.3)$$

Dabei ist der mittlere Ankunftsabstand (Erwartungswert,vgl. Anhang A1)

$$E(T_A) = 1/\lambda$$

bzw. die mittlere Bedienungszeit

$$E(T_H) = 1/\varepsilon \qquad (1.4)$$

Eine für die verkehrstheoretische Berechnung sehr angenehme Eigenschaft liegt darin,daß derart verteilte Zufallsvariable kein "Gedächtnis" besitzen,sondern daß die restliche Dauer einer negativ-exponentiell verteilten zufälligen Dauer unabhängig von der seitherigen Dauer t_o ist.Es gilt also z.B. bei derartig verteilten Bedienungszeiten

$$P(T_H > t_o+t \mid T_H > t_o) = \frac{P(T_H > t_o+t, T_H > t_o)}{P(T_H > t_o)} = \frac{e^{-\varepsilon(t_o+t)}}{e^{-\varepsilon t_o}} = P(T_H > t) \qquad (1.5)$$

Man sagt auch,solche Zufallsvariable besitzen die "Markoff--Eigenschaft",da ihre zukünftige Dauer nur vom gegenwärtigen Zustand (d.h.vom Bestehen) abhängt.

Zur verkehrstheoretischen Berechnung müssen die VF der Ankunftsabstände und der Bedienungszeiten (Ankunfts- und Bedienungsprozesse) bekannt sein. Jedoch genügt in vielen Fällen auch eine Angabe über Mittelwert und Varianz σ^2 der VF, d.h. ihr erstes und zweites Moment (vgl.Anhang A1).

Ein wichtiger Parameter einer VF ist der Varianzkoeffizient

$$C = \frac{\sigma}{m_1} \qquad (1.6)$$

mit m_1 als erstem Moment (Erwartungswert) der VF.Er kann als normierte Standardabweichung betrachtet werden. Speziell bei der negativ-exponentiellen VF gilt C=1,bei VF mit kleinerer Varianz (sog. hypoexponentielle VF bzw. Varianzen) gilt $0 \leq C < 1$, bei VF mit größerer Varianz (sog. hyperexponentielle VF) ist $C > 1$.

Im Anhang A1 sind neben vollständigen Definitionen der VF-Größen die hier verwendeten VF aufgezählt und einheitlich mit allen Momenten dargestellt.

Als Kennung für VF-Typen und Systeme wird in dieser Arbeit die allgemein benutzte Notation von KENDALL [38] verwendet,in der z.B. M|G|n ein einstufiges Wartesystem mit negativ-exponentiell verteilten Ankunftsabständen,beliebig verteilten Bedienungszeiten und n BE bedeutet (siehe Abkürzungsverzeichnis Punkte 5,6).

Es sei an dieser Stelle noch eine andere übliche Kennzeichnung der Verkehrsarten genannt,bei der z.B. reiner Zufallsverkehr (unendliche Zahl von Quellen,negativ-exponentiell verteilte Ankunftsabstände und Bedienungszeiten) als Zufallsverkehr 1.Art (ZV1) bezeichnet wird.

Bei der verkehrstheoretischen Behandlung stochastischer Systeme werden im allgemeinen bei der Modellierung noch zusätzliche Unabhängigkeitsannahmen getroffen,die als Voraussetzungen in das verkehrstheoretische Modell eingehen.

Dies sind zunächst Annahmen über die Unabhängigkeit von Zufallsvariablen(Ankunftsabstände bzw.-zeitpunkte,Bedienungszeiten) aufeinanderfolgender Anforderungen,so daß beide Prozesse "rekurrent" sind,d.h. daß bei Start einer neuen,nach einer bestimmten VF verteilten Zeitdauer die Vergangenheit unerheblich ist,der Prozeß also quasi neu zu laufen beginnt.(Dies ist z.B. nicht der Fall,wenn der"Ankunftsprozess" dadurch zustande kommt, daß aus peripheren Einheiten jeweils nur maximal eine begrenzte

Zahl von Anforderungen mit einem meist starren Takt in die Bedienungsstufe übernommen werden,vgl. z.B.|53|.)
Weiterhin wird angenommen,daß die Bedienungszeit einer Anforderung nicht von ihrem Ankunftsabstand zur Vorgängerin abhängt. Es sei auch noch die allg. übliche Annahme erwähnt,daß die Bedienungszeiten nicht durch nachfolgende Anforderungen beeinflußt werden,was bei technischen Systemen praktisch immer erfüllt ist.

● Als Betriebsparameter gelten alle Angaben bezüglich der Strategien zur Vergabe der Betriebsmittel "Warteplätze" und"BE" (vgl. etwa |51,52|).Dazu gehören die Auswahlstrategien

- innerhalb einer Warteschlange (Abfertigungsdisziplin)
- innerhalb einer Gruppe von BE (Absuchmodus)
- bei Zusammenführung und Verzweigung des Verkehrs.

Interessiert man sich in obigem einstufigen,vollkommen erreichbaren Wartesystem nur für die Gesamtzahl der momentan belegten BE,so ist der Absuchmodus ohne Einfluß.Es verbleibt als wichtigste Betriebsstrategie die Abfertigungsdisziplin der wartenden Anforderungen in der Warteschlange.Diese Rangordnung der Anforderungen für das Nachrücken in eine freigewordene BE kann evtl. nach (externen) Prioritäten gemäß der Art oder Herkunft erfolgen (vgl. z.B. |29,45|) und dann innerhalb einer Klasse von Anforderungen gleicher Priorität gekennzeichnet sein durch

- den Ankunftszeitpunkt (Ankunftsreihenfolge) (FIFO,LIFO)
- reinen Zufall (RANDOM)
- die geforderte Bedienungszeit (SJF)
- die bislang erhaltene Bedienungszeit bei Zulassung von Unterbrechungen.

Die wichtigste Abfertigungsdisziplin ist die mit Abfertigung in Ankunftsreihenfolge (FIFO,first-in first-out),bei der die am längsten wartende Anforderung einer Klasse ausgewählt wird. Dagegen wird die inverse Abfertigungsreihenfolge (LIFO,last-in first-out) dann angewendet,wenn die zuletzt angekommene Anforderung am wichtigsten ist (z.B. bei Meßdaten).In technischen Systemen lässt sich die Abfertigungsstrategie auch häufig durch eine zufällige Auswahl (RANDOM) beschreiben.
Darüber hinaus gibt es auch solche Disziplinen,die sich an der geforderten Gesamt- oder Restbedienungszeit orientieren(SJF, shortest-job-first etc.).

Im Rahmen dieser Arbeit wird zur Anschauung bevorzugt die FIFO-Strategie herangezogen;solange jedoch nur Mittelwerte von Wartezeiten etc. betrachtet werden und nicht deren VF, sind diese z.B. unabhängig von den Abfertigungsdisziplinen FIFO,LIFO,RANDOM (conservation law,vgl.|12|).Diese Unabhängigkeit gilt jedoch nur solange,wie sich die Abfertigungsdisziplin nicht ex- oder implizit an den Bedienungszeiten orientiert.

1.2.2 Charakteristische Verkehrsgrößen

Zur Beschreibung der Verkehrsgüte in Systemen mit statistisch schwankender Zahl von Belegungen werden sog. charakteristische Verkehrsgrößen benutzt.In vielen Fällen interessiert das Verhalten des Systems bei Stationarität,wenn es sich nach einer gewissen Zeit im eingeschwungenen Zustand befindet (statistisches Gleichgewicht).Dann sind die Wahrscheinlichkeiten für die Zustände des Systems (Zahl der Anforderungen im System,etc) von der Zeit und den Anfangsbedingungen unabhängig und gleich den absoluten Zustandswahrscheinlichkeiten des Systems.Entsteht bei verschiedenen Anfangsbedingungen stets derselbe stationäre Zustand,so liegt Ergodizität des Systems vor.
Bezeichnet man die mittlere Zahl der pro Zeiteinheit eintreffenden Anforderungen (Ankunftsrate) mit λ,so ist bei einer mittleren Bedienungszeit $E(T_H) = h$ der sog. angebotene Verkehr (Angebot)

$$A \stackrel{def}{=} \lambda \cdot h \qquad (1.7)$$

Bei einem reinen Wartesystem muß, um stationäre Bedingungen zu erhalten (d.h. die mittlere Warteschlangenlänge nicht stetig anwachsen zu lassen),bei vollkommener Erreichbarkeit das Angebot A kleiner sein als die Zahl n der BE.Für die Ausnutzung ϱ einer BE muß also gelten:

$$\varrho = \frac{A}{n} < 1 \qquad \text{für Stationarität (1.8)}$$

Wird mit X die zufällige Zahl aller Anforderungen im System, d.h. in BE und Wartespeicher,bezeichnet,so kann man die stationären Zustandswahrscheinlichkeiten des Systems definieren:

$$p(x) \stackrel{def}{=} P(X=x) \qquad (1.9)$$

Bei dem betrachteten vollkommen erreichbaren Wartesystem gelingt die Zustandsbeschreibung mit einer Komponente, im allgemeinen jedoch (bei unvollkommener Erreichbarkeit,mehrstufigen Anordnungen...) sind mehrere Komponenten zur eindeutigen Charakterisierung eines Systemzustandes notwendig (Zustandsvektor).
Besitzt das System keinen oder nur einen endlich großen Speicher,so geht mit einer gewissen Wahrscheinlichkeit B (Verlustwahrscheinlichkeit) eine Anforderung "verloren". Solange der Ankunftsprozeß ein sog. Poisson-Prozeß ist (Ankunftsabstände negativ-exponentiell verteilt) trifft in jedem Augenblick gleichwahrscheinlich eine Anforderung ein und die Verlustwahrscheinlichkeit ist identisch mit der absoluten Wahrscheinlichkeit,daß das System voll belegt ist:

$$B = p(n+s) \qquad \text{bei Poisson-Ankunftsprozeß} \quad (1.10)$$

Damit beträgt die Rate der erfolgreichen Anforderungen (durchgesetzte Rate)

$$\lambda_Y = \lambda \cdot (1-B) \qquad (1.11)$$

und die Belastung Y aller BE

$$Y = A \cdot (1-B) = \lambda_Y \cdot h \ , \qquad (1.12)$$

die auch als Erwartungswert der Zahl belegter BE dargestellt werden kann:

$$Y = \sum_{x=0}^{n+s} \min(x,n) \cdot p(x) \qquad (1.13)$$

Zu Ehren des dänischen Verkehrstheoretikers A.K.ERLANG (vgl.|3|) dient als Maßeinheit für diese fiktive bzw. tatsächliche mittlere Gleichzeitigkeit (A bzw. Y) bestehender Belegungen die dimensionslose Größe 1 Erlang,abgekürzt Erl .

Bei Systemen mit Wartemöglichkeit können Anforderungen auf freiwerdende BE warten.Wichtigste Kenngröße hierfür ist der Erwartungswert $E(T_W)$ der zufälligen Wartezeit T_W von Anforderungen,der die mittlere Wartezeit bezogen auf alle Anforderungen angibt.Darüber hinaus interessiert die Wahrscheinlichkeit W,überhaupt warten zu müssen (Wartewahrscheinlichkeit) und,falls eine Anforderung warten muß,ihre mittlere Wartezeit t_W (mittlere Wartezeit der Wartenden)

$$t_w = E(T_w \mid T_w > 0) \qquad (1.14)$$

Als Beziehung zwischen diesen Größen gilt

$$E(T_w) = W \cdot t_w \qquad (1.15)$$

Eine vollständige Information über die Wartezeiten ist jedoch nur durch ihre VF gegeben.Diese gibt an,mit welcher Wahrscheinlichkeit eine zufällige Wartezeit höchstens gleich einer Zeit t ist,bzw. bei der komplementären VF diese überschreitet:

$$W(>t) = P(T_w > t)$$

mit (1.16)

$$W(>0) = W$$

und

$$E(T_w) = \int_{t=0}^{\infty} t \cdot f_w(t)\,dt = \int_{t=0}^{\infty} W(>t)\,dt \qquad (1.17)$$

wobei $f_W(t)$ die Dichtefunktion der Wartezeiten bezogen auf alle Anforderungen darstellt.

Für die mittlere Zahl Ω wartender Anforderungen (Länge der Warteschlange,Wartebelastung) gilt allgemein die Beziehung

$$\Omega = \lambda \cdot E(T_w) \ , \qquad (1.18)$$

sie kann aber auch direkt aus den Zustandswahrscheinlichkeiten abgeleitet werden:

$$\Omega = \sum_{z=1}^{s} z \cdot p(n+z) \qquad (1.19)$$

Die Zeit,die eine Anforderung im System verbringt,sei mit Durchlaufszeit T_F bezeichnet,sie setzt sich aus der Warte- und Bedienungszeit zusammen:

$$T_F = T_W + T_H \qquad \text{(Zufallsvariable)} \quad (1.20)$$

Sind Warte- und Bedienungszeit einer Anforderung voneinander

unabhängig,so ergibt sich die VF der Durchlaufszeit aus der Faltung der VF (vgl.|9,15|)

$$F(\leq t) = P(T_F \leq t) = W(\leq t) * H(\leq t) \qquad (1.21)$$

Nach einem Satz aus der Wahrscheinlichkeitstheorie (siehe z.B. |9|) gilt für die mittlere Durchlaufszeit

$$E(T_F) = E(T_W) + E(T_H), \qquad (1.22)$$

auch wenn Wartezeit und Bedienungszeit einer Anforderung voneinander abhängig sein sollten (z.B. bei SJF,shortest job first).

Die mittlere Zahl von Anforderungen im gesamten System E(X) setzt sich aus der Belastung Y und der Wartebelastung Ω zusammen

$$E(X) = \Omega + Y = \lambda \cdot E(T_F) \qquad (1.23)$$

und kann auch direkt aus den Zustandswahrscheinlichkeiten bestimmt werden:

$$E(X) = \sum_{x=0}^{n+s} x \cdot p(x) \qquad (1.24)$$

Rein formal kann man nun die charakteristischen Verkehrsgrößen in 2 Gruppen einteilen:

- Verkehrsgrößen mit direkter Aussage über <u>das System</u>
- Verkehrsgrößen mit direkter Aussage über <u>das Schicksal von Anforderungen</u>

Zu den Aussagen über das System gehören z.B. die Zustandswahrscheinlichkeiten,die mittlere Zahl von Anforderungen im System,in den BE,im Wartespeicher,sowie Aussagen über Zeiten ununterbrochener Belegungen (busy period) oder Freizeiten (idle period).
Größen wie z.B. die Wartezeitverteilungsfunktion (mit mittl. Wartezeiten,Wartewahrscheinlichkeit) oder die VF der Durchlaufszeiten sind primär Aussagen über das Schicksal von Anforderungen,aber natürlich mit dem System selbst verknüpft. Bei einem reinen Verlustsystem z.B. muß man im allg. zwischen der (absoluten) Wahrscheinlichkeit,daß das Bündel voll belegt ist und der Verlustwahrscheinlichkeit unterscheiden; beide

Aussagen sind nur identisch für einen Poisson-Ankunftsprozess.

Als wichtige Beziehung zwischen einigen Verkehrsgrößen aus diesen beiden Gruppen kann eine in der Literatur unter dem Namen "LITTLEs formula" |41| oft benutzte,wichtige Beziehung gelten.Diese,unter sehr allgemeinen Bedingungen geltende (fast triviale) Beziehung sagt,daß (bei Stationarität) die mittlere Zahl von Anforderungen in einem beliebigen Teilsystem gleich dem Produkt aus der Rate der Anforderungen,die in dieses Teilsystem gelangen und der mittleren Aufenthaltszeit aller dieser Anforderungen in diesem Teilsystem ist.
Verschiedene Gleichungen dieses Abschnitts sind Beispiele hierfür,wobei in der Gleichung für

Y	(1.12)	die Gesamtheit der BE
Ω	(1.18)	der Wartespeicher
E(X)	(1.23)	das gesamte System betrachtet wurde.

Über die in diesem Abschnitt aufgeführten Verkehrscharakteristika hinaus gibt es noch weitere z.B. bei Systemen mit unterbrechenden Prioritäten die Unterbrechungswahrscheinlichkeit, die mittl. Zahl von Unterbrechungen etc. .Bei diesen komplexeren Verkehren gelten obige Definitionen dann sinngemäß auch individuell pro Prioritätsklasse,ebenso wie bei komplexeren Strukturen auch für Teilsysteme.

1.2.3 Berechnungsmethoden

Bevor einzelne Berechnungsmethoden genannt werden,ist es sinnvoll,zunächst die verwandten Begriffe "Prozeß" und "Kette" gegeneinander abzugrenzen (vgl. Bild 1.2).
Unter einem Prozeß X(t) versteht man den Verlauf des (hier diskreten) Zustandes X eines Systems über einem kontinuierlichen Parameter,meist der Zeit t,während bei einer Kette nur Aussagen über die Zustände zu bestimmten diskreten Zeitpunkten des Prozesses gemacht werden (vgl. z.B.|2| u.a.).

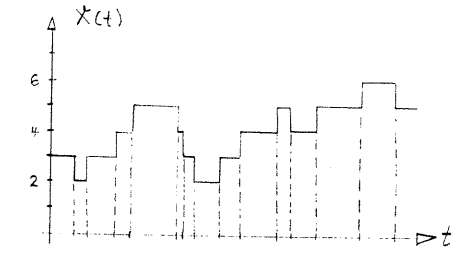
Je nach den getroffenen Annahmen über die statistischen Eigenschaften des Verkehrs (VF der Ankunfts-und Bedienungsprozesse),kann man zwischen

- Markoff'schen Prozessen

und

- Nicht-Markoff'schen Prozessen

unterscheiden.



Prozeß X(t)
(Zustandsverlauf
über der Zeit t)

..,3,2,3,4,5,4,3,2,3,4,5,4,5,6,5,..

Kette
(Folge von z.B.
allen Zuständen)

.. 2, 4,3,2, 4, 5,..

Eingebettete Kette
(Folge von Zuständen
zu bestimmten Zeit-
punkten)

Bild 1.2 Zur Definition von
"Prozeß" und "Kette"

Rein Markoff'sche Prozesse besitzen zu jedem Zeitpunkt die sog. "Markoff-Eigenschaft",d.h. der zukünftige Prozeßverlauf hängt nur vom momentanen Zustand des Prozesses ab (auch Geburts- und Sterbe-Prozeß genannt).
Bei Nicht-Markoff'schen Prozessen hängt der zukünftige Prozeßverlauf von der Vorgeschichte ab,und man versucht, diesen "Prozeß mit Nachwirkung" auf einen Prozeß mit Markoff-Eigenschaft zurückzuführen,indem man entweder den Prozeß nur zu solchen Zeitpunkten betrachtet,in denen die Markoff-Eigenschaft erfüllt ist, oder indem man durch Zusatzinformationen in der Zustandsbeschreibung einen Markoff-Prozeß erzeugt.
Ist nun der vergangene Prozeßverlauf zu bestimmten Zeitpunkten (Regenerationszeitpunkten) unerheblich,beginnt also der Prozeß quasi neu zu laufen,so spricht man auch von einem rekurrenten Prozeß oder Erneuerungsprozeß.

● Markoff'sche Prozesse sind,von der Methode her, relativ einfach zu berechnen,deshalb wurden diese in der Literatur eingehend behandelt |13,16,19,42|.

Durch Aufstellen des sog. "Kolmogoroff schen Gleichungssystems" (System linearer Differentialgleichungen) und den Übergang zu stationären Bedingungen (Nullsetzen der Differentialquotienten) entsteht ein homogenes lineares Gleichungssystem für die stationären Zustandswahrscheinlichkeiten, in denen die Koeffizienten i.allg. einfach angebbare Übergangswahrscheinlichkeiten bzw. -dichten darstellen.Durch Ersetzen einer beliebigen Gleichung durch die Normierungsbedingung

$$\sum_{x=0}^{\infty+s} p(x) = 1$$

entsteht ein inhomogenes Gleichungssystem mit einer eindeutigen Lösung für die stationären Zustandswahrscheinlichkeiten, aus denen dann relativ einfach weitere Verkehrsgrößen bestimmt werden können. Benutzt man sog.erzeugende Funktionen (vgl. etwa |9|),so können solche Verkehrsgrößen auch oft direkt aus dem Gleichungssystem bestimmt werden,ohne daß die Zustandswahrscheinlichkeiten selbst bekannt sein müssen.

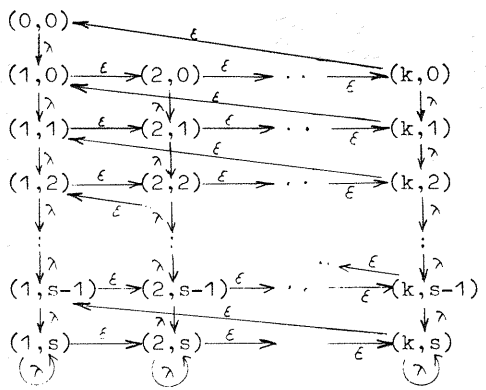
● Zur Berechnung Nicht-Markoff scher Prozesse werden bevorzugt folgende Methoden angewendet:

- Phasenmethode nach Erlang
- Methode der eingebetteten Markoff-Kette
- Methode der supplementären Variablen
- Lindley'sche Integralmethode

● Bei der Phasenmethode nach ERLANG |3| werden die Ankunftsabstände und Bedienungszeiten durch negativ-exponentiell-verteilte fiktive (Teil-)Phasenzeiten dargestellt und in der Zustandsbeschreibung der Index der momentan belegten Phase zusätzlich angegeben.
Bild 1.3 zeigt als Beispiel ein solches Zustandsdiagramm mit der entsprechenden Wahl der Zustandsbeschreibung für das Warte-Verlustsystem mit Erlang-k-verteilten Bedienungszeiten (System $M|E_k|1$-s).
Setzt man z.B. die Stationaritätsbedingung für Zustände an, die keine "Randzustände" sind, so erhält man (vgl.Bild1.3)

$$(\lambda+\varepsilon)\,p(j,z) = \lambda\cdot p(j,z-1)+\varepsilon\cdot p(j-1,z) \qquad j=2..k;\ z=1..s-1$$

Fügt man noch die Gleichungen für die restlichen Zustände hinzu, erhält man (mit der Normierungsbedingung) ein inhomogenes Gleichungssystem wie bei rein Markoff'schen Voraussetzungen.



Zustandsbeschreibung

p(j,z)

$j = \begin{cases} 1..k & \text{Nummer der Bedienungsphase} \\ 0 & \text{BE frei} \end{cases}$

z= 0..s Zahl wartender Anforderungen

Bild 1.3 Zustandsdiagramm für System $M|E_k|1\text{-}s$ ($E(T_H) = k \cdot \frac{1}{\varepsilon}$)

● Bei der Methode der eingebetteten Markoff-Kette,die auf KENDALL |38| zurückgeht,wird der Nicht-Markoff'sche Prozeß nur zu ganz bestimmten Zeitpunkten betrachtet.Diese sog. Regenerationszeitpunkte sind je nach System verschieden,z.B. im einstufigen Wartesystem M|G|1 sind dies die Zeitpunkte, zu denen eine Anforderung das System verlässt (vgl.Bild 1.2); dann ist die Zahl der Anforderungen im System nur abhängig vom Zustand des Systems bei Abgang der letzten Anforderung und der zufälligen Zahl der inzwischen angekommenen Anforderungen.Als Ergebnis der Berechnung der Markoff-Kette erhält man die Zustandswahrscheinlichkeiten des Systems an den Regenerationszeitpunkten,die aber nicht unbedingt mit den absoluten Zustandswahrscheinlichkeiten zu beliebigen Zeitpunkten übereinstimmen. Eine solche eingebettete Markoff-Kette wird auch in dieser Arbeit in 6.3.2 betrachtet.

● Bei der Methode der supplementären Variablen wird das System wieder zu allen Zeitpunkten betrachtet,jedoch zur Erzeugung eines Prozesses mit Markoff-Charakter ein zusätzlicher kontinuierlicher Parameter benötigt,der die Bedeutung einer verstrichenen Zeitdauer besitzt,z.B. im System M|G|1 die seitherige Bedienungszeit der Anforderung,die gerade bedient wird (Bedienungsalter),vgl COX |32|.

● Bei der Integralmethode nach LINDLEY |40| werden Beziehungen zwischen den Wartezeiten aufeinanderfolgender Anforderungen aufgestellt und hieraus eine Integralgleichung abgeleitet,die es

gestattet,die Wartezeit-VF im stationären Falle zu ermitteln. Die Lösung für diese Integralgleichung kann jedoch oft nur unter bestimmten Voraussetzungen angegeben werden|44|.

1.2.4 Literaturhinweise und benutzte Ergebnisse

Die Entwicklung der Warteschlangentheorie ging 1909 von den ersten Berechnungen von ERLANG |3| für Vermittlungssysteme aus,wurde in den dreißiger Jahren von POLLACZEK und KHINTCHINE (vgl.z.B. |5|) fortgesetzt und führte in den fünfziger und sechziger Jahren zu einer großen Zahl von Veröffentlichungen auf diesem Gebiet.Eine 1957 veröffentlichte Bibliographie über Warteschlangentheorie |8| enthält bereits ca. 750 Titel; eine Übersicht über behandelte Probleme mit mehr als 900 Literaturangaben wurde von SAATY |16|gegeben und 7 Jahre später "wehklagend" eine Flut von nahezu 2000 Literaturstellen registriert|17|.
Die stürmische Entwicklung der Rechnertechnik und die damit verbundene Möglichkeit der numerischen Auswertung und der Notwendigkeit von verkehrstheoretischen Lösungen |47,49| haben bis heute zu einer starken Ausweitung der zugehörigen Literatur geführt und zu einer Vielzahl von Ergebnissen.
In dieser Arbeit werden verschiedene Ergebnisse über einstufige Wartesysteme benutzt.Dies sind u.a. Zustandswahrscheinlichkeiten und Wartezeit-VF für das System M|M|n,mittlere Wartezeiten im System M|G|1,sowie Größen im Warte-Verlustsystem M|M|1-s (Notationen vgl. Abkürzungsverzeichnis Punkt 6).
Ebenfalls benötigte Ergebnisse für die mittleren Wartezeiten aller Anforderungen im System GI|G|1 wurden im Rahmen einer anderen Arbeit |39| approximativ ermittelt.

Von besonderem Interesse sind in diesem Zusammenhang auch solche Veröffentlichungen über einstufige Wartesysteme, die sich mit dem Ausgangsprozeß dieser Systeme befassen (vgl. 2.2.1), da dieser bei serieller Anordnung mehrerer solcher reiner Wartesysteme mit dem Eingangsprozeß in die jeweils nachfolgende Stufe identisch ist.

1.3 Mehrstufige Wartesysteme

Nachdem in 1.2 eine kurze Übersicht über einstufige Modelle gegeben wurde, können nun mehrstufige Wartesysteme behandelt werden, die durch serielle Anordnung mehrerer Einzelstufen gebildet werden. In diesem Abschnitt sollen zunächst die Funktionsweisen mehrstufiger Wartesysteme geschildert und ihre charakteristischen Verkehrsgrößen beschrieben werden, um dann auf besondere Fragestellungen bei der Dimensionierung solcher Systeme einzugehen.

1.3.1 Allgemeines

Bei vielen technischen Systemen bildet der Ausgangsprozeß einer Stufe den Eingangsprozeß in eine nachfolgende Stufe; d.h. die Anforderungen benötigen zur vollständigen Bearbeitung mehrere Bearbeitungsphasen in verschiedenen seriell angeordneten Bedienungsstufen (vgl. Bild 1.4).

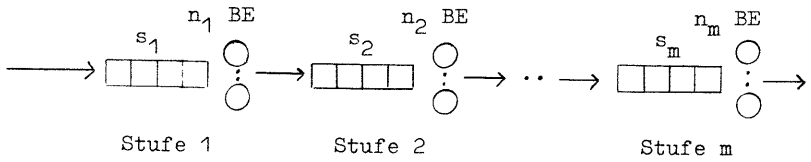


Bild 1.4 System mit seriellem Warten

Dabei kann eine Anforderung in jeder Stufe gezwungen sein, auf das Freiwerden einer BE zu warten, deshalb spricht man von "mehrstufigen Wartesystemen" oder von "Systemen mit seriellem Warten", in der englisch-sprachigen Literatur von "Queues in Series", "Queues in Tandem" oder (nicht ganz treffend) von "Queues with Phase-Type Service", auch von "Production Line".

Eine große Rolle bei diesen Systemen spielt die Größe der Zwischenspeicher (Stufe 2 bis m). Ist z.B. bei einem 2-stufigen Wartesystem die Größe s_2 des Zwischenspeichers in der 2. Stufe unbegrenzt, so kann jede Anforderung nach Beendigung ihrer Bedienung in Stufe 1 sofort in Stufe 2 gelangen und die freigewordene BE in Stufe 1 ihren Service an einer evtl. schon angekommenen

Nachfolgeanforderung fortsetzen. Ist dagegen in der 2. Stufe gar kein oder nur ein endlich großer Speicher vorgesehen ($0 \leqslant s_2 < \infty$), so ist es wegen der statistisch schwankenden Bedienungszeiten möglich, daß eine Anforderung bei Beendigung ihrer eigentlichen Bedienungszeit in Stufe 1 keinen freien Warteplatz (bzw. bei $s_2=0$ keine freie BE) in Stufe 2 vorfindet und dadurch eine BE "blindbelegt". Je länger die Zeit dauert, bis in der nächsten Stufe ein Platz frei wird, desto länger bleibt diese BE für nachfolgende Anforderungen gesperrt. In der Literatur über mehrstufige Wartesysteme wird dies allg. als "Blockierung" bezeichnet. Dies bedeutet hier nicht, daß eine eintreffende Anforderung abgewiesen werden muß, wie bei Verlustsystemen, sondern daß diese quasi außerhalb der Stufe warten kann.
Im wesentlichen unterscheidet man bezüglich der Struktur also zwischen mehrstufigen Wartesystemen

- ohne Blockierung (unendlich große Zwischenspeicher)
- mit Blockierung (endlich große bzw. ohne Zwischenspeicher)

Verluste können hier nur vor der 1. Stufe auftreten, sofern diese keinen unbegrenzt großen Wartespeicher besitzt.
Den Durchlauf von Anforderungen durch das gesamte System kann man z.B. anhand eines Zeitdiagrammes für den Aufenthalt einer Anforderung in einer Stufe i des Systems verfolgen (vgl. Bild 1.5).

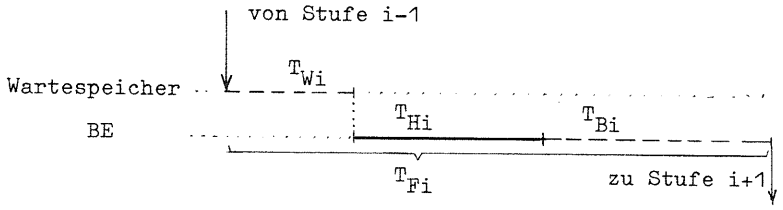


Bild 1.5 Zeitdiagramm einer Anforderung in Stufe i

Die Zeit T_{Fi} einer Anforderung in Stufe i (Durchlaufszeit, flowtime) ist eine Zufallsvariable und setzt sich zusammen aus

- der Wartezeit T_{Wi} im Wartespeicher
- der Bedienungszeit T_{Hi} und
- der Blockierzeit T_{Bi} in einer BE

$$T_{Fi} = T_{Wi} + T_{Hi} + T_{Bi} \qquad (1.25)$$

Dabei sind die Wartezeit und die Blockierzeit in Bild 1.5 nur gestrichelt eingezeichnet,da diese vom momentanen Zustand des Systems abhängen,also auch gleich 0 sein können.
Die gesamte Durchlaufszeit T_F einer Anforderung durch das mehrstufige Wartesystem beträgt

$$T_F = \sum_{i=1}^{m} T_{Fi} \qquad (1.26)$$

die gesamte Verzögerungszeit T_{WB} (Warten und/oder Blockierung) ist

$$T_{WB} = \sum_{i=1}^{m} (T_{Wi} + T_{Bi}) \qquad (1.27)$$

Für Systeme ohne Blockierung ist diese identisch mit der Gesamtwartezeit T_W.

Bei der Berechnung wird durchweg angenommen,daß die Bedienungszeiten T_{Hi} (i=1..m) der gleichen Anforderung in den aufeinanderfolgenden Stufen des Systems voneinander unabhängig sind (vgl. z.B. "Independence Assumption" |12|),wodurch eine Abhängigkeit zwischen dem Ankunftsabstand einer Anforderung und ihrer geforderten Bedienungszeit in der Stufe 2,3,... vermieden wird.

1.3.2 Charakteristische Verkehrsgrößen

● Da sich mehrstufige Wartesysteme aus Einzelstufen zusammensetzen,kann man zunächst einmal viele Verkehrsgrößen wie in einstufigen Systemen definieren (vgl. 1.2.2);diese erhalten nur einen zusätzlichen Index, der die betreffende Stufe angibt. Für die bei Einzelstufen nicht auftretende Blockierung seien folgende Verkehrsgrößen genannt:

$$\left.\begin{array}{ll} P(T_{Bi}>0) \overset{def}{=} p_{Bi} & \text{Blockierwahrscheinlichkeit} \\ E(T_{Bi}) & \text{mittlere Blockierzeit bezogen auf alle Anforderungen} \\ E(T_{Bi} \mid T_{Bi}>0) \overset{def}{=} t_{Bi} & \text{mittlere Blockierzeit der Blockierten} \end{array}\right\} (1.28)$$

Außerdem gilt

$$E(T_{Bi}) = p_{Bi} \cdot t_{Bi} \qquad (1.29)$$

und für die mittlere Zahl Y_{Bi} gleichzeitig blockierter BE (Blockierbelastung)

$$Y_{Bi} = \lambda \cdot E(T_{Bi}) \qquad (1.30)$$

Dabei ist λ die gesamte Ankunftsrate vor Stufe 1;d.h. $E(T_{Bi})$ bezieht sich auf alle Anforderungen,die in Stufe 1 ankommen. Treten Verluste vor der 1. Stufe auf,so ist es sinnvoll,eine mittlere Blockierzeit bezogen auf alle erfolgreichen Anforderungen anzugeben, deren Rate λ_Y natürlich aus Gründen der Stationarität für alle Stufen gleich ist (solange keine Verzweigungen auftreten):

$$\lambda_{Yi} = \lambda_Y = \lambda \cdot (1-B) \quad \text{für alle } i = 1..m \qquad (1.31)$$

Für diese mittlere Blockierzeit aller Erfolgreichen

$$E(T_{Bi} \mid \text{erfolgreich}) \overset{def}{=} E(T_{Bi})_Y \qquad (1.32)$$

gilt nun sinngemäß

$$Y_{Bi} = \lambda_Y \cdot E(T_{Bi})_Y \qquad (1.33)$$

In der letzten Stufe m tritt keine Blockierung der BE auf.

● Neben diesen Größen für Einzelstufen gibt es auch solche, die sich auf das Gesamtsystem beziehen,diese sollen keinen Index erhalten:

$p(x)$	Wahrscheinlichkeit für x Anforderungen im Gesamtsystem
$E(X)$	Mittlere Zahl von Anforderungen im gesamten System
$E(T_F)$	Mittlere Gesamtdurchlaufszeit
$p(x_1,x_2,..x_m)$	Wahrscheinlichkeit für Zustandsmuster $\{x_1,x_2,..x_m\}$

usw.

● Eine weitere Art von Verkehrsgrößen sind solche,die man nur durch Verfolgen des Schicksals einer bestimmten Anforderung im Gesamtsystem erhalten kann.Dieses Gesamtschicksal setzt sich aus ihren Teilschicksalen in den verschiedenen Stufen zusammen.
Hierzu zählt z.B. die Gesamtwartewahrscheinlichkeit W einer Anforderung.Dies ist die Wahrscheinlichkeit,daß eine

beliebige Anforderung irgendwo im Gesamtsystem warten muß:

$$P(T_W>0) \overset{def}{=} W \qquad \text{wobei } T_W=\sum_{i=1}^{m} T_{Wi} \tag{1.34}$$

Damit verknüpft ist die mittlere Gesamtwartezeit t_W der Wartenden, die sich aus

$$W\cdot t_W = E(T_W) = \sum_{i=1}^{m} E(T_{Wi}) \tag{1.35}$$

bestimmen läßt.

Entsprechende Größen kann man auch für die Verzögerungszeiten (Summe aus Warte- und Blockierzeiten) definieren:

$$P(T_{WB}>0) \overset{def}{=} p_{WB} \tag{1.36}$$

bzw.

$$p_{WB}\cdot t_{WB} = E(T_{WB}) \tag{1.37}$$

Mit diesen Verkehrsgrößen läßt sich nun die Verkehrsgüte in den verschiedenen Stufen und im Gesamtsystem angeben. Bevor jedoch auf Fragestellungen bei der Dimensionierung solcher Systeme eingegangen wird, soll zunächst das Verhalten des Systems im wichtigen Grenzfalle des maximalen Durchsatzes genauer beschrieben werden.

1.3.3 Systemverhalten bei maximalem Durchsatz

Bei mehrstufigen Wartesystemen mit durchweg unendlich großen Zwischenspeichern wird die im stationären Grenzfall gerade noch verarbeitbare Rate λ_{max} von Anforderungen durch die langsamste Stufe bestimmt, die den "Flaschenhals" des Systems darstellt:

$$\lambda_{max} \overset{def}{=} \lambda_{Ymax} = \min(\mu_1,..\mu_m) \tag{1.38}$$

wobei $\mu_i = n_i/h_i$ die maximal verarbeitbare Rate einer Einzelstufe darstellt (Enderate bei voll belegter Stufe). Die mittlere Durchlaufszeit $E(T_{Fi})$ durch diese "Flaschenhalsstufe" geht gegen unendlich, während sie für andere Stufen endlich bleibt (verschiedene Stufen vorausgesetzt).

Bei mehrstufigen Wartesystemen mit nur endlich großen bzw. nicht vorhandenen Zwischenspeichern ist -abgesehen von der letzten Stufe- in jeder Stufe Blockierung möglich. Dabei tritt nicht nur ein Rückstau über eine Stufe auf, sondern es kann z.B. eine langsame letzte Stufe oder eine zufällig sehr große Bedienungszeit in der letzten Stufe die Ursache für einen Rückstau bis in die 1. Stufe sein. Durch die endlich großen Zwischenspeicher wirkt also jeder Engpaß des Systems auf die 1. Stufe zurück. Deshalb ist die maximal verarbeitbare Rate λ_{max} dann erreicht, wenn alle BE der 1. Stufe immer belegt sind (Blockierung eingeschlossen):

$$Y_1 + Y_{B1} = n_1 \qquad \text{bei } \lambda_Y = \lambda_{max} \tag{1.39}$$

Diese Betriebsweise der Grenzbelastung des Systems bzw. der 1. Stufe kann auf verschiedene Weise erreicht werden: bei begrenztem Speicher in der 1. Stufe ($0 \leq s_1 < \infty$) muß die Ankunftsrate λ gegen unendlich gehen und damit die Verlustwahrscheinlichkeit gegen 1; bei $s_1 \to \infty$ wäre bei $\lambda > \lambda_{max}$ die Warteschlange in Stufe 1 instationär.

Diese maximale Rate ist bei statistisch schwankenden Bedienungszeiten wegen Blockierung immer kleiner als der Wert nach (1.38) und von allen Stufen abhängig.
Diese Rate wird bei Systemen ohne Verzweigung durch alle Stufen durchgesetzt, es gilt somit

$$\lambda_{max} = \frac{Y_1}{h_1} = \frac{Y_2}{h_2} = \ldots = \frac{Y_m}{h_m} \qquad \text{bei } \lambda_Y = \lambda_{max} \tag{1.40}$$

Ihr Kehrwert $1/\lambda_{max}$ kann generell als mittlere Zeit zwischen zwei aufeinanderfolgenden Zu- bzw. Abgängen einer beliebigen Stufe betrachtet werden, also insbesondere auch als mittlerer Abgangsabstand hinter der (nicht blockierten) letzten Stufe oder auch als mittlerer Zugangsabstand zu Stufe 1.
Da die Wartezeit einer erfolgreichen Anforderung im Wartespeicher der Stufe 1 von der willkürlich gewählten Speichergröße s_1 abhängt bzw. bei $s_1 \to \infty$ über alle Grenzen wächst, ist es sinnvoll, in diesem Grenzfalle das System ohne diesen Wartespeicher zu betrachten, so daß diese Wartezeit (gewissermaßen außerhalb des betrachteten Systems) nicht relevant ist. Hierzu kann

man sich vorstellen,daß die 1.Stufe jedesmal bei Freiwerden einer BE sofort eine neue Anforderung aus einem unendlich großen Reservoir holt (vgl.spezielles Symbol in Bild 1.6).

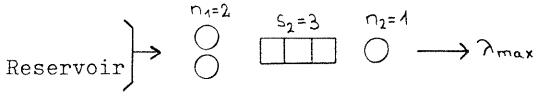


Bild 1.6 2-stufiges Strukturbeispiel für maximalen Durchsatz

Diese Betriebsweise ist z.B. besonders interessant bei Fertigungsstraßen, wo am Ausgangspunkt beliebig viele Werkstücke bereitliegen.
Die maximale Durchsatzrate kann als wichtigste Kenngröße eines Systems mit Blockierung gelten,da sie mit derjenigen Ankunftsrate identisch ist,bei der in einem solchen reinen Wartesystem die mittlere Gesamtdurchlaufszeit (und damit auch die mittlere Zahl von Anforderungen im Gesamtsystem) eine Asymptote besitzt. Eine Kenntnis des Verhältnisses der Ankunftsrate zur maximal verarbeitbaren Rate ist in vielen Fällen (z.B. zur ersten Abschätzung der Verzögerungszeiten) von großem Vorteil.

1.3.4 Verkehrstheoretische Fragestellungen

Bei der verkehrstheoretischen Berechnung und der Simulation solcher serieller Systeme stehen bei gegebener Struktur und gegebenem Verkehr zunächst einmal Fragen der Auslastung der Systemkomponenten im Vordergrund. Hierzu gehören u.a. die Ausnutzung der BE,die Belegung des Wartespeichers (Warteschlangenlängen) sowie eventuelle Blockierbelastungen.Dabei kann das System bei einem gewissen oder -ebenfalls zu bestimmenden- maximalen Durchsatz betrachtet werden.
Eng mit diesen Größen verknüpft sind Aussagen bezüglich der Warte-,Blockier- und Durchlaufszeiten von Anforderungen,sowie entsprechender Wahrscheinlichkeiten für Warten und/oder Blockierung,wie auch des Verlustes (Verkehrsgüte).Diese Größen können auf Einzelstufen bezogen sein,aber auch auf das Gesamtsystem. Im letzteren Fall muß das Schicksal einer Anforderung in den aufeinanderfolgenden Stufen verfolgt werden.Dies ist besonders bei Systemen mit Blockierung notwendig,weil dort die Teilschicksale derselben Anforderung pro Stufe voneinander stärker abhängig sind.

Bei geforderter Verkehrsgüte einer Einzelstufe oder des Gesamtsystems wird eine Systemstruktur gesucht,die diese Bedingungen mit möglichst wenig Aufwand erfüllt. Hierzu zählen die Fragen nach der Zahl der BE oder der Warteplätze,um z.B. die Forderungen bezüglich des Verlusts bzw. des Wartens und/oder der Blockierung zu erfüllen.Aber auch Aussagen bezüglich der notwendigen Schnelligkeit von BE sind oft Ziel der Untersuchungen. Von besonderem Interesse ist die Frage,wie groß bei gegebenen Schwankungen der Bedienungszeiten die Zwischenpuffer sein müssen, um einen bestimmten maximalen Durchsatz zu erreichen. Eine solche Fragestellung kann z.B. mit den in Kapitel 5 dieser Arbeit geschilderten Ergebnissen für 2-stufige Single-Server-Systeme allgemein beantwortet werden.
Es kann aber auch diejenige Reihenfolge der Stufen interessieren,die bei gegebener Durchsatzrate die kleinste mittlere Durchlaufszeit ergibt.

1.4 Anwendungsbereiche mehrstufiger Wartesysteme

In diesem Abschnitt sollen einige der vielen Anwendungsbereiche serieller Wartesysteme genannt und mit typischen Beispielen belegt werden.Abgesehen von Systemen des täglichen Lebens (Kraftfahrzeuginspektion,Verkaufsschalter etc.) treten diese als Systeme bzw. Teilsysteme hauptsächlich innerhalb von Rechnern,in Vermittlungssystemen und auch bei Fertigungsprozessen auf.

1.4.1 Serielles Warten in Rechnern

Elektronische Datenverarbeitungsanlagen sind in ihrer Struktur und Betriebsweise sehr komplex,so daß bei diesen entweder nur solche Modelle untersucht werden können,die das Geschehen nur global beschreiben,oder daß einzelne Teilkonfigurationen herausgegriffen werden müssen.Hiermit sollen die verschiedenen Möglichkeiten angesprochen sein, Modelle des Ablaufgeschehens in Rechnern zu bilden auf

- Programmebene oder
- Befehlsebene (Programm- oder Steuerbefehle)

bzw. in einer Zwischenebene,die z.B. bei Rechnern mit"Paging" durch die Seiten eines Programms dargestellt wird oder z.B. allgemein durch Teilaufgaben von Programmen,die ohne Unterbrechung in der Zentraleinheit (CPU) abgearbeitet werden könnten.

Zu all diesen Beispielen ist zu bemerken, daß die Abarbeitung von Anforderungen (z.B.Zeit für Ausführung e.Befehls) für diese selbst determiniert ablaufen,jedoch für das System als Ganzes nach Zeitpunkt und Dauer den Charakter eines Zufallsereignisses besitzt.

Bild 1.7a zeigt ein einfaches Rechnermodell zur Abarbeitung von Programmen oder Teilaufgaben in einem Rechner mit zweistufiger Speicherhierarchie (Arbeitsspeicher ASP,Hintergrundspeicher HSP),bei dem ein gleichzeitiges unabhängiges Arbeiten des Rechnerkerns (CPU) und des Schnellkanals (SK) möglich ist (Multiprogramming).

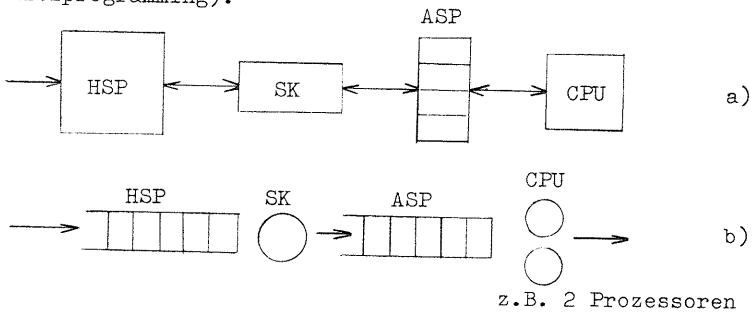


Bild 1.7 a) Einfaches Rechnermodell zur Abarbeitung von Programmen

b) Warteschlangenmodell (Rücktransporte unberücksichtigt)

Es wird angenommen, daß alle Programme auf dem Hintergrundspeicher (Band,Platte,Trommel) stehen bzw. dort eintreffen und nacheinander zur Bearbeitung in den Arbeitsspeicher transportiert werden. Würde dieses Modell durch Einbeziehen des Rücktransports verfeinert werden,so entstünde ein sog.zyklisches Wartesystem (vgl.2.4.1).

Die Verkehrseigenschaften der Systeme und damit die verkehrstheoretischen Fragestellungen sind eng mit dem jeweiligen Anwendungsfall verknüpft;als wichtigste Größe bei reinem Stapelbetrieb gilt der maximale Durchsatz,während bei Rechnern mit Teilnehmerbetrieb oder Prozeßrechnern gefordert wird,daß eine bestimmte Antwortzeit nur mit einer kleinen Wahrscheinlichkeit überschritten wird.

Als Beispiel für die Anwendung serieller Warteschlangenmodelle in Rechnern auf Befehlsebene sei eine Befehls-Pipeline genannt, die in Bild 1.8 dargestellt ist.

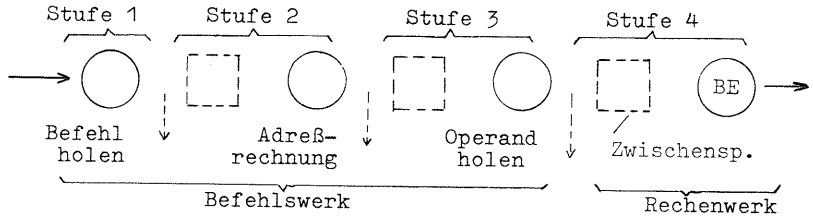


Bild 1.8 Befehls-Pipeline

Dabei ist die Bearbeitung von Befehlen in einem Prozessor so organisiert,daß mehrere Befehle gleichzeitig in einzelnen autonomen Unterwerken aufbereitet bzw.ausgeführt werden.Da verschiedene Befehle einzelne Unterwerke verschieden lang belegen,kommt es zu Verzögerungszeiten (Blockierzeiten auf BE, Wartezeiten bei evtl. vorhandenen Zwischenpuffern).Dies macht sich insgesamt in einer Verringerung der maximalen Durchsatzrate von Befehlen bemerkbar,die für die Leistungsfähigkeit einer Rechenanlage von entscheidender Bedeutung ist.

Für den allgemeinen Fall (vgl.|105|) kann man etwa noch Verzweige- oder Ausstiegswahrscheinlichkeiten hinter einzelnen Stufen vorsehen (z.B.für Sprungbefehle ohne/mit Adressumrechnung,Transportbefehle).Da die ganze Befehls-Pipeline mit einem gemeinsamen Takt betrieben wird,können die Bedienungszeiten der Befehle in den einzelnen Stufen durch eine Mehrpunkt-VF beschrieben werden,bei der alle Bedienungszeiten ganzzahlige Vielfache dieser Taktzeit sind (vgl. Anhang A1.7 und Anhang A2 ,wo solche VF u.a. betrachtet werden).

Ähnliche serielle Modelle treten in mehr oder weniger abgewandelter Form als Zugriffsmodelle für periphere Datenspeicher auf.

Einen Einblick in die Problemstellungen und die Anwendung der Warteschlangentheorie in Rechnern findet sich z.B. in |46-53, 55|.

1.4.2 Serielles Warten in Vermittlungs-bzw.Übertragungssystemen

Neuere Vermittlungssysteme für Daten- und Fernsprechverkehr werden oft durch Rechner gesteuert.Deshalb ergeben sich auch hier prinzipiell die gleichen Problemstellungen und Anwendungsfälle wie in 1.4.1 .
Darüber hinaus können beispielsweise Datennetze,die nach dem sog. "Message- oder Packet-Switching-Prinzip" arbeiten,durch komplexe Wartesysteme beschrieben werden.Eine Teilkonfiguration,nämlich die eines Vermittlungsknotens,ist in Bild 1.9 dargestellt.

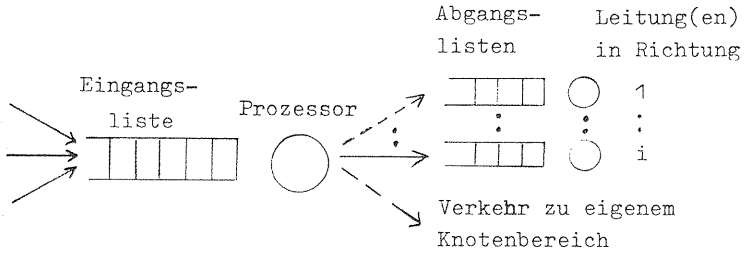


Bild 1.9 Modell eines Datenvermittlungsknotens

Die aus den verschiedenen Richtungen eintreffenden "Messages" oder "Packets" werden in den Speicher des Knotens eingeschrieben und in die Liste der wartenden Anforderungen (Warteschlange) eingetragen zur Auswertung ihrer Kopfinformation (Zielrichtung etc) durch den Prozessor.Dann erfolgt die Einweisung (Eintrag in eine Warteliste) in die gewünschte Richtung zum nächsten Knoten.Deren Leitung wird dann nach eventuellem weiteren Warten für die Dauer der Übertragungszeit belegt.
Ein solches Modell wird hier in Kapitel 3 behandelt;u.a. mit zusätzlicher sog. alternativer Leitweglenkung von HERZOG |11|.

In Nachrichtenübertragungssystemen wird häufig eine Zeitmultiplextechnik angewandt,bei der die einzelnen Nachrichten verschiedener Herkunft von einem sog. Multiplexer in Blöcke (Gruppen von Zeichen) unterteilt werden und dann nacheinander über dieselbe Übertragungsleitung gesendet werden.
In vielen Anwendungsfällen (z.B. bei Teilnehmerkonsolen,die an einen Rechner angeschlossen sind) tritt ein sog."Burst-Betrieb" auf,d.h. die Teilnehmer geben ihre Nachrichten stoßweise ab

(z.B. am Bildschirm eingetastete Zeile eines Programms),vgl. Bild 1.10.

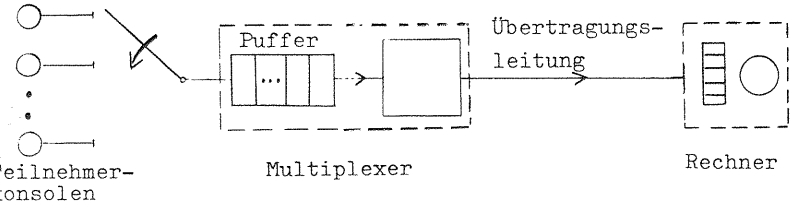


Bild 1.10 Datenübertragung zu einem Rechner(im asynchronen Zeitmultiplex)

Deshalb ist es hier nicht unbedingt sinnvoll,bei 1 Übertragungsleitung jeder Konsole eine periodisch wiederkehrende Zeitdauer fester Länge (Zeitschlitz) zuzuteilen ("synchrones Zeitmultiplex"), sondern diese Zuteilung asynchron vorzunehmen.
Hierbei ist die Zuteilung der Zeitschlitze nicht starr,sondern erfolgt nach Bedarf,weshalb jeder Block mit zusätzlichen Adressinformationen versehen werden muß (vgl. Bild 1.11).

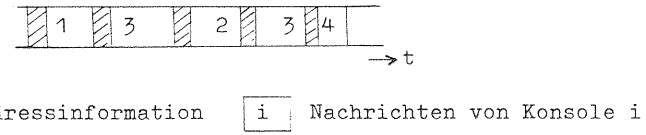


Bild 1.11 Asynchrones Zeitmultiplex

Die zusätzliche Adressinformation wird durch den Multiplexer erzeugt,dessen Bearbeitungsdauer als konstant angenommen werden darf.Wegen der statistisch auftretenden Verkehrsspitzen muß der Multiplexer einen Puffer enthalten,wie auch Warten im Rechner erlaubt ist.
Man hat hier also ein zweistufiges Wartesystem (Multiplexer mit Übertragungsleitung,Rechner),bei dem die Nachrichten bereits außerhalb des Rechners warten müssen, erhält aber dafür wegen eines "Glättungseffektes" (vgl. 2.2.1) des Multiplexers kleinere Wartezeiten im Rechner selbst.

2 ÜBERSICHT ÜBER DIE IN DER LITERATUR BEHANDELTEN SYSTEME, METHODEN UND ERGEBNISSE

Als Hinführung zu den eigentlichen und neuen Untersuchungen dieser Arbeit in den Kapiteln 3-6 und als Motivation für die dort behandelten Systeme wie auch der verwendeten Berechnungsmethoden wird in diesem Kapitel ein Überblick über die in der umfangreichen Literatur der mehrstufigen Wartesysteme vorhandenen analytischen Untersuchungen gegeben und kurz auf die verwendeten Berechnungsmethoden eingegangen (vgl. Bild 2.0).

Es werden die wichtigsten Ergebnisse genannt, bzw. einige, auf die in Kap. 3-6 Bezug genommen wird, ausführlicher behandelt. Die Übersicht in diesem Kapitel ist systemorientiert aufgebaut und enthält alle dem Verfasser bekannten diesbezüglichen Veröffentlichungen.

2.1 Einführung

Serielle Wartesysteme können nach ihrer Systemstruktur oder nach ihren Verkehrsarten klassifiziert werden.
Bezüglich der Systemstruktur können sie nach der Größe der Zwischenspeicher (vgl. 1.3.1) eingeteilt werden in Systeme

- ohne Blockierung (unbegrenzte Zwischenspeicher)
- mit Blockierung (keine bzw. endlich große Zwischenspeicher)

bezüglich der Verkehrsart in Systeme mit

- rein Markoff'schen Voraussetzungen
- Nicht-Markoff'schen Voraussetzungen

Bei Systemen ohne Blockierung kann jede Anforderung nach Ablauf ihrer regulären Bedienungszeit sofort in die nächste Stufe gelangen, weil dort ein unbegrenzter Wartespeicher zur Verfügung steht. Der Ausgangsprozeß einer solchen Stufe ist deshalb identisch mit demjenigen einer Einzelstufe mit gleichem Eingangs- und Bedienungsprozeß wie in der betrachteten mehrstufigen Anordnung. Deshalb sind Aussagen über Ausgangsprozesse von einstufigen Wartesystemen von besonderem Interesse. Auf diese wird in 2.2.1 eingegangen, wo auch gezeigt wird, daß diese Vorgehensweise wegen möglicher Abhängigkeiten nur unter bestimmten Voraussetzungen möglich ist.

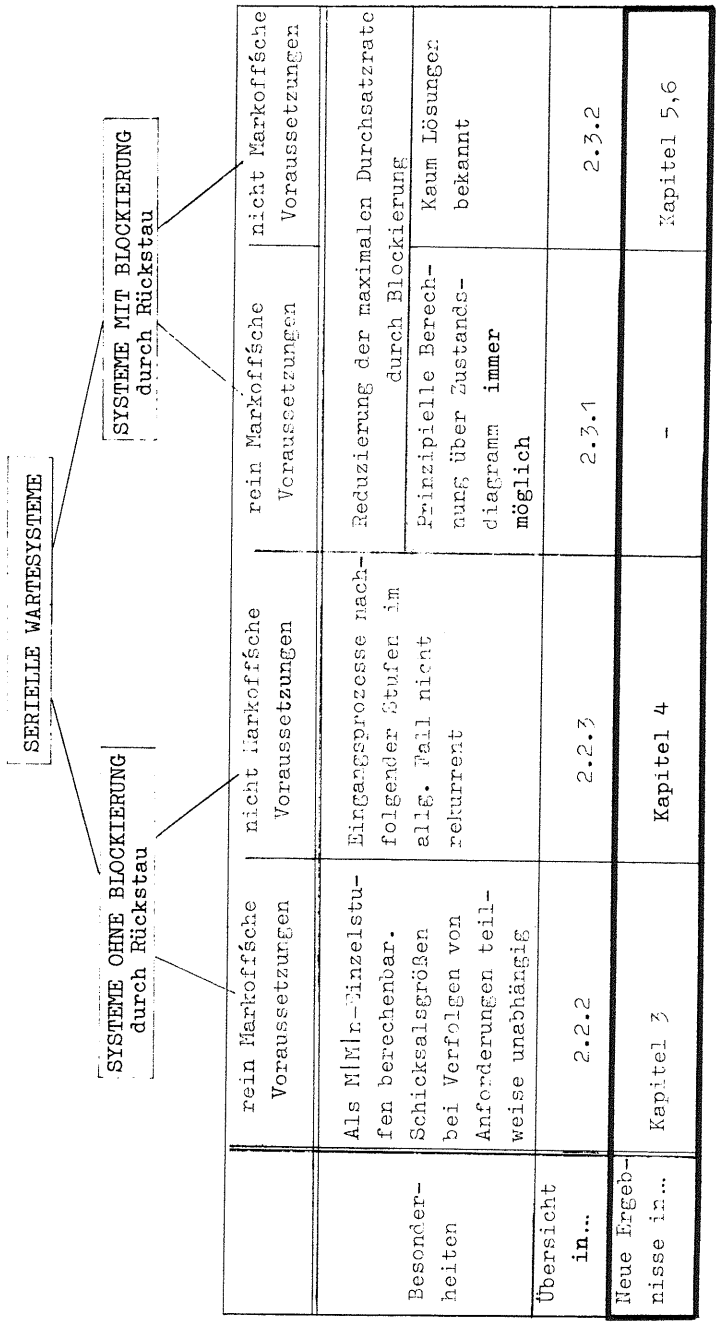


Bild 2.0 Klassifizierung von seriellen Wartesystemen und Kapitelhinweise

● Bei Systemen mit Blockierung kann aufgrund der beschränkten Zahl von Warteplätzen in einem Zwischenspeicher eine Blindbelegung von Bedienungseinheiten (BE) auftreten,wenn die nächste Stufe momentan voll belegt ist.Dieser Blockiereffekt von BE in einer Stufe i bedeutet eine (zustandsabhängige) Vergrößerung der gesamten Belegungszeit einer Anforderung durch die Blockierzeit T_{Bi}.Im Falle $s_{i+1}>0$ folgt der Blockierzeit T_{Bi} eine Wartezeit $T_{Wi+1}>0$.
Somit kann eine blockierte BE in Stufe i als momentane Verlängerung des Wartespeichers der Stufe i+1 betrachtet werden,wobei aber die Ankunftsrate in Stufe i+1 momentan reduziert bzw bei $n_i=1$ zu Null wird.Eine blockierbare BE hat also neben ihrer eigentliche Funktion als Serviceeinheit auch die eines Warteplatzes für die nächste Stufe.Deshalb können in einem System mit seriellem Warten die Zwischenspeicher durch Gruppen von BE mit Bedienungszeit ≡0 dargestellt werden (vgl. Bild 2.1).

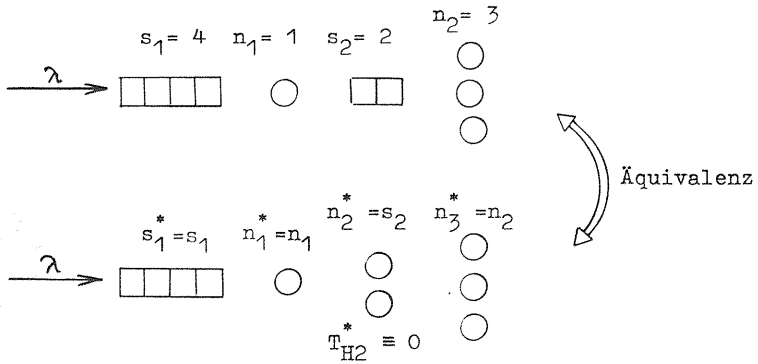


Bild 2.1 Beispiel für Abbildung eines Systems mit Zwsp.auf ein äquivalentes System ohne Zwischenspeicher

Auf diese Äquivalenz wird z.B. in |56| hingewiesen.
Zur Blockierung ist noch allgemein zu sagen,daß es auch Systeme gibt,in denen scheinbar keine Blockierung auftritt,die aber durch ein System mit Blockierung voll beschrieben werden können.

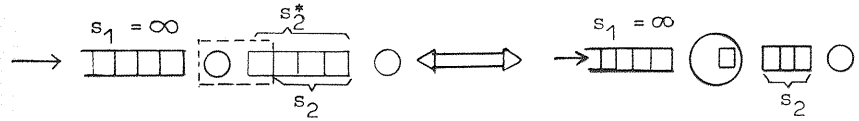


Bild 2.2 Zweistufiges System mit sichergestelltem Warteplatz in 2.Stufe

Bild 2.2 zeigt ein solches System,bei dem -entgegegen der hier üblichen Betrachtungsweise- die Bedienungszeit in Stufe 1 nur dann beginnen darf,wenn in der 2.Stufe noch mindestens einer von den s_2^* Warteplätzen frei ist,also rein formal keine Blockierung auftritt. Betrachtet man jedoch den zur Bedienung notwendigen letzten Warteplatz in Stufe 2 als Bestandteil der BE in Stufe 1, verbleiben $s_2=s_2^*-1$ Warteplätze und die seitherige Wartezeit auf dem letzten Warteplatz entspricht genau der Blockierzeit in den hier betrachteten Modellen.

● Bei Systemen mit rein Markoff'schen Voraussetzungen sind der Ankunftsprozeß und die Bedienungsprozesse durch eine negativ-exponentielle VF gekennzeichnet.Dadurch ist es im Prinzip einfach,das Zustandsdiagramm für die stationären Zustandswahrscheinlichkeiten anzugeben (vgl. 1.2.3) und durch analytische oder numerische Verfahren zu lösen.

● Bei Systemen mit Nicht-Markoff'schen Voraussetzungen müssen z.B. die in 1.2.3 geschilderten Methoden für Nicht-Markoff'sche Prozesse sinngemäß angewendet werden.Welche der dort angeführten Methoden zum Ziel führt,hängt von dem untersuchten System ab;es gelingt z.B. nicht immer, eine eingebettete Markoff-Kette anzugeben |89|,d.h. dieses Verfahren anzuwenden.

In der Literatur über Wartesysteme in Serie findet man zunächst relativ viele Veröffentlichungen über rein Markoff'sche Systeme ohne Blockierung,von denen einige Ergebnisse auch durch Anwendung eines Theorems von BURKE über den Ausgangsprozeß eines Systems M|M|n gewonnen werden können.
Rein Markoff'sche Systeme mit Blockierung wurden hauptsächlich für den Grenzfall des maximalen Durchsatzes (vgl. 1.3.3) betrachtet und dieser meist bei relativ kleiner Stufenzahl als Funktion der Systemstruktur und der mittleren Bedienungszeiten bestimmt.Dabei wurden für mehrere BE je Stufe nur 2-stufige Systeme betrachtet,bei "Single-Server-Systemen" auch 3 Stufen, bzw. bei identischen Single-Server-Stufen auch eine höhere Stufenzahl.

Bei Nicht-Markoff'schen Systemen ohne Blockierung wurden fast nur Single-Server-Stufen behandelt,jedoch für mehr als 2 Stufen nur Stabilitätsbedingungen abgeleitet.

Bei Nicht-Markoff'schen Systemen mit Blockierung wurden vornehmlich 2 seriell angeordnete BE ohne Zwischenspeicher untersucht, bei beliebigen VF der Bedienungszeiten in beiden Stufen. Systeme mit nur konstanten Bedienungszeiten nehmen wegen ihres weitgehend deterministischen Charakters eine Sonderstellung ein und wurden deshalb auch bei allgemeinerer Struktur in der Literatur behandelt.
Der rasch ansteigende Aufwand zur exakten Lösung dieser Probleme bei beliebigen VF der Bedienungszeiten kann anhand eines 2-stufigen Single-Server-Systems von NEUTS |89| (1968) mit Zwischenspeicher ($s_2>0$) angedeutet werden. Dieses besitzt jedoch noch eine negativ-exponentielle VF der Bedienungszeiten in Stufe 2. Mit der in vorliegender Arbeit verwendeten Kurzdarstellung für serielle Wartesysteme (ähnlich der von KENDALL |39|, vgl. Abkürzungsverzeichnis Punkt 6) kann es dargestellt werden als System

$$\underbrace{M \xrightarrow{\infty} G|1}_{\text{Stufe 1}} \quad \underbrace{\xrightarrow{s_2} M|1}_{\text{Stufe 2}}$$

Dieses relativ einfache System benötigt jedoch zu einer exakten Lösung, wie sie von NEUTS hergeleitet wurde, einen solchen mathematischen Lösungsaufwand, daß BURKE |62|, auf den wichtige Ergebnisse über serielle Warteschlangen zurückgehen, hierüber schreibt: "The results are so complicated as to put their usefulness into question".
Vor diesem Hintergrund sind insbesondere auch die Näherungsmethoden mit ihren Ergebnissen zu sehen, die in vorliegender Arbeit entwickelt werden (Kap.4-6).
Die in der Literatur angewandten Methoden sind vorwiegend exakt und können als sinngemäße Anwendung bzw. Erweiterung der Methoden angesehen werden, über die in Abschnitt 1.2.3 im Zusammenhang mit einstufigen Wartesystemen berichtet wurde.

2.2 Systeme ohne Blockierung

Durch unendlich große Zwischenspeicher wird bei mehrstufigen Wartesystemen erreicht, daß jede in einer Stufe bearbeitete Anforderung sofort in die nächste Stufe gelangen kann. Deshalb sind die Einzelstufen rückwirkungsfrei und der Eingangsprozeß in eine Stufe i ist gleich dem ungestörten Ausgangsprozeß der

Stufe i-1. Gelingt es, diesen Ausgangsprozeß zu bestimmen, so ist das System auf Einzelstufen zurückgeführt und es können Ergebnisse von einstufigen Systemen angewandt werden. Dabei muß jedoch leider gefordert werden, daß nicht nur die VF für die Ausgangsabstände bekannt ist, sondern, -um überhaupt bekannte Ergebnisse von einstufigen Wartesystemen anwenden zu können- daß auch aufeinanderfolgende Ausgangs- bzw. Ankunftsabstände voneinander unabhängig sind (d.h. ein rekurrenter Prozeß vorliegt). Deshalb wird in 2.2.1 zunächst auf Ausgangsprozesse bei einstufigen Systemen eingegangen.

2.2.1 Ausgangsprozesse bei einstufigen Wartesystemen

Je mehr serielle Systeme und Netze Gegenstand von Untersuchungen wurden, desto interessanter und notwendiger waren Aussagen über Ausgangsprozesse bei einstufigen Systemen. Den Auftakt hierzu bildete der Beweis, daß der Ausgangsprozeß einer einstufigen Warteanordnung M|M|n ein Poissonprozeß ist (Theorem von BURKE |30|, 1956). Er bewies mit Hilfe von Differentialgleichungen für bedingte Abgangsabstands-VF, daß die Zahl der Abgänge in einem willkürlich gewählten Zeitintervall die gleiche ist wie für die Ankünfte und daß die Zeitintervalle zwischen zwei aufeinanderfolgenden Abgängen unabhängig voneinander und negativ-exponentiell verteilt sind, d.h. einen Poissonprozeß bilden (vgl. hierzu auch COHEN |64|).
Zusammen mit der Tatsache, daß bei wahrscheinlichkeitsmäßiger Verzweigung eines Poissonprozesses in mehrere Teilprozesse wieder Poissonprozesse entstehen und daß bei Zusammenführung von unabhängigen Poissonprozessen wieder ein Poissonprozeß entsteht, stellt dieses keineswegs selbstverständliche Ergebnis die Grundlage zur Berechnung einiger Grundstrukturen dar (serielle Anordnungen, Netze ohne Rückkopplung..).
Das gleiche Theorem wurde von REICH |94| mit Hilfe sog. reversibler Markoff-Prozesse bewiesen und gezeigt, daß eine Erweiterung von BURKE's Theorem auf allgemeinere VF nicht möglich ist. Hierzu bewies REICH

a) daß der Ausgangsprozeß aus einer Stufe $E_k|E_k|n$ nur für k=1 (negativ-exponentiell) und (trivialerweise) auch für $k\to\infty$ (konstante VF) ein E_k-Prozeß ist.

und b) daß der Ausgangsprozeß einer Stufe M|VF$_H$|1 nur dann ein Poissonprozeß sein kann, wenn die VF der Bedienungszeiten VF$_H$ eine negativ-exponentielle VF ist mit Mittelwert h ≧ 0.

Bei weiteren Versuchen der Verallgemeinerung bzw. Abgrenzung von BURKE's Theorem zeigt z.B. FINCH |35|, daß beim Ausgangsprozeß aus einem einem einstufigen Warte-Verlustsystem M|M|1-s bei endlich großer Speichergröße s aufeinanderfolgende Abgangsabstände nicht mehr voneinander unabhängig sind.
Dagegen gelang es BOES |28| zu zeigen, daß bei dem einstufigen Warte-Verlustsystem M|M|n-s der Ausgangsprozeß ein Poissonprozeß ist, sofern man die Verlustanforderungen als am Ausgangsprozeß beteiligt auffaßt. Dies gilt ebenfalls bei Zulassung von Verzichten zum Zeitpunkt des Eintreffens einer Anforderung oder nach einer negativ-exponentiell verteilten Geduldszeit (balking, reneging).
MAKINO |83| bestimmte u.a. die momenterzeugende Funktion (vgl. z.B. |9|) der Abgangsabstände des Ausgangsprozesses (departure process) im Wartesystem M|G|1 und daraus die zugehörige Varianz bzw. den Varianzkoeffizienten C_D

$$C_D^2 = 1 - A^2 \cdot (1 - C_H^2) \qquad (2.1)$$

Dabei ist $C_H = \sigma_H / h$ der Varianzkoeffizient der Bedienungszeit-VF und $A = \lambda \cdot h$ das Angebot.

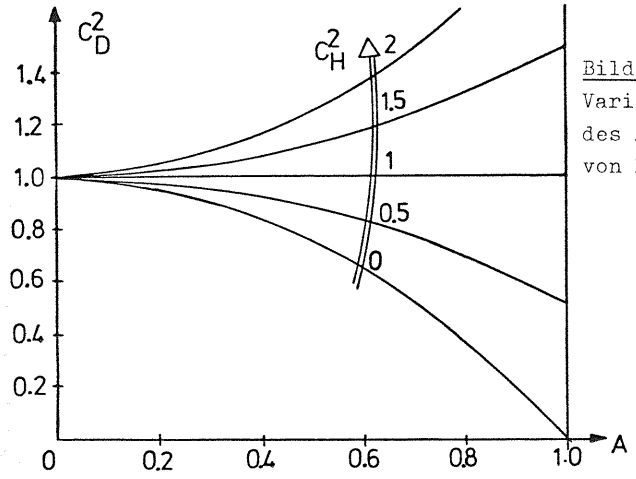


Bild 2.3
Varianzkoeffizient C_D des Ausgangsprozesses von M|G|1

Wie aus Bild 2.3 ersichtlich, liegt der Varianzkoeffizient C_D des Ausgangsprozesses bei allen Angeboten A zwischen dem des Ankunftsprozesses C_A (=1, Poisson) und dem des Bedienungsprozesses C_H, z.B.

$$1 \geqslant C_D \geqslant C_H \quad \text{bei hypoexponentiellen VF } (C_H < 1).$$

Dazu muß aber bemerkt werden, daß im obigen Falle im allgemeinen Abhängigkeiten zwischen aufeinanderfolgenden Abgangsabständen bestehen, die (wie in 4.2.2 gezeigt werden wird) bei seriellen Systemen einen spürbaren Einfluß haben können. Solche Korrelationen wurden z.B. von COX, vgl. bei |96|, für M|G|1, von JENKINS |36| für M|E$_k$|1 sowie von CHANG |31| und DALEY |33| für das System GI|G|1 untersucht.
DALEY zeigte, daß im allgemeinen System GI|G|1 die Varianz des Ausgangsprozesses größer ist als die des Bedienungsprozesses

$$\mathrm{Var}(T_D) \geqslant \mathrm{Var}(T_H), \qquad (2.2)$$

was für den spezielleren Fall M|G|1 auch aus den Ergebnissen von MAKINO hervorgeht. Ein Vergleich mit der Varianz des Ankunftsprozesses ergab eine beliebige Relation

$$\mathrm{Var}(T_D) \gtrless \mathrm{Var}(T_A). \qquad (2.3)$$

Darüber hinaus gibt es noch weitere Veröffentlichungen über Ausgangsprozesse, die aber im Rahmen dieser Arbeit von geringerem Interesse sind, wie z.B. Ausgangsprozesse bei ∞ vielen BE |43| oder VF für die Zeit bis zum Abgang einer gewissen Zahl von Anforderungen etc.
Weiterführende Literatur hierüber findet sich z.B. in REICH |96| und BURKE |62|.

2.2.2 Systeme mit rein Markoff'schen Voraussetzungen

Wartesysteme in Serie ohne Blockierung mit rein Markoff'schen Voraussetzungen wurden in der Literatur häufig behandelt. Dabei verlagerte sich -nach Aufstellung von BURKE's Theorem - das Interesse von der Bestimmung der Zustandswahrscheinlichkeiten der Einzelstufen und des Gesamtsystems auf die Untersuchung des Gesamtdurchlaufs von Anforderungen (vgl. 1.3.2) und die dabei auftretenden Fragen der Abhängigkeiten bzw. Unabhängigkeiten von Teilschicksalen (Wartezeiten, Durchlaufszeiten etc).

2.2.2.1 Berechnung der Einzelstufen und des Gesamtsystems

Zu den ersten Veröffentlichungen über serielle Wartesysteme zählt O'BRIEN |58|, der 1954 für das 2-stufige Single-Server System (vgl. Abkürzungsverzeichnis Punkt 6)

$$M \xrightarrow{\infty} M|1 \xrightarrow{\infty} M|1$$

die mittleren Warteschlangenlängen und Durchlaufszeiten in beiden Stufen, sowie die mittlere Gesamtdurchlaufszeit bestimmte. Dabei nahm er heuristisch einen Poissonverkehr als Eingangsprozeß für Stufe 2 an (ist erfüllt, vgl. 2.2.1).
R.R.P.JACKSON |75| betrachtete zunächst das gleiche System und bewies durch Ansetzen und Lösen der Zustandsgleichungen, daß die Zustandswahrscheinlichkeiten der Einzelstufen zu gleichen Zeitpunkten voneinander unabhängig sind, was er in |76| auf Multiserver-Systeme beliebiger Stufenzahl

$$M \xrightarrow{\infty} M|n_1 \xrightarrow{\infty} M|n_2 \xrightarrow{\infty} \cdots \xrightarrow{\infty} M|n_m$$

erweiterte (Theorem von R.R.P.JACKSON, 1956):

$$p(x_1, x_2, \ldots x_m) = \prod_{i=1}^{m} p_i(x_i) \qquad (2.4)$$

Dies gilt nach BURKE |62| auch bei beliebiger VF in der letzten Stufe. Wegen dieser Unabhängigkeit kann man z.B. bei solchen Systemen die Gesamtzahl der Anforderungen im System durch (diskrete) Faltung gewinnen.
Ebenfalls 1956 wurde das Theorem von BURKE |30| über den Ausgangsprozeß einer Stufe M|M|n aufgestellt (vgl. 2.2.1), mit Hilfe dessen die Einzelstufen direkt berechnet werden können.
Diese beiden Ergebnisse (von BURKE und R.R.P. JACKSON) sind grundlegend für die Berechnung serieller Systeme, die aus unabhängigen Einzelstufen mit Markoff-Eigenschaft bestehen.

2.2.2.2 Berechnung des Gesamtschicksals von Anforderungen

Durch die Betrachtung einer beliebigen festen Anforderung während ihres Durchlaufs durch ein mehrstufiges Wartesystem können (vgl.1.3.2) Aussagen über Verkehrsgrößen gemacht werden, die sich auf das Gesamtschicksal von Anforderungen beziehen. Das Gesamtschicksal einer Anforderung im System (z.B. ihre Gesamtwartezeit T_W) setzt sich zusammen aus ihren Teilschicksalen in den einzelnen Stufen (z.B. Wartezeiten T_{Wi}). Sind diese voneinander unabhängig, so kann man das Gesamtschicksal (VF, Erwartungswerte..) durch Faltung etc. erhalten.
Auf die Frage des Zusammenhangs dieser Teilschicksale wird hier ausführlicher eingegangen, da bezüglich dieser in Kapitel 3 neue Ergebnisse hergeleitet werden und die Kenntnis bestehender Ergebnisse hierüber eine Voraussetzung bildet.
Betrachtet man ein z.B. 2-stufiges System

$$M \xrightarrow{\infty} M|n_1 \xrightarrow{\infty} M|n_2$$

mit FIFO-Abfertigungsdisziplin in jeder Stufe und als Schicksalsgrößen (Zufallsvariable) einer Anforderung ihre Wartezeit T_{Wi}, ihre Bedienungszeit T_{Hi} und ihre Durchlaufszeit T_{Fi} in Stufe i (i=1,2), so kann man zwischen diesen Größen derselben Anforderung bestimmte Beziehungen (Abhängigkeit, Unabhängigkeit) angeben. Bild 2.4 stellt eine graphische Übersicht dar über einige Beziehungen zwischen verschiedenen Schicksalsgrößen derselben Anforderung im obigen System.

Neben den oben beschriebenen Zufallsvariablen von Zeitgrößen sind noch die Zahl X_{Ai} der bei Ankunft in Stufe i angetroffenen Anforderungen, die Zahl X_{D1} der in Stufe 1 zurückgelassenen Anforderungen aufgeführt, sowie die vergangene Ausgangsfolge Prev.dep.seq. (Previous departure sequence) der 1.Stufe bis zum Abgang der betrachteten Anforderung aus Stufe 1. Die Doppelpfeile zwischen zwei Größen geben an, ob diese beiden Größen (separat betrachtet) voneinander abhängig sind oder nicht. Außerdem sind sie mit Literaturreferenzen versehen.

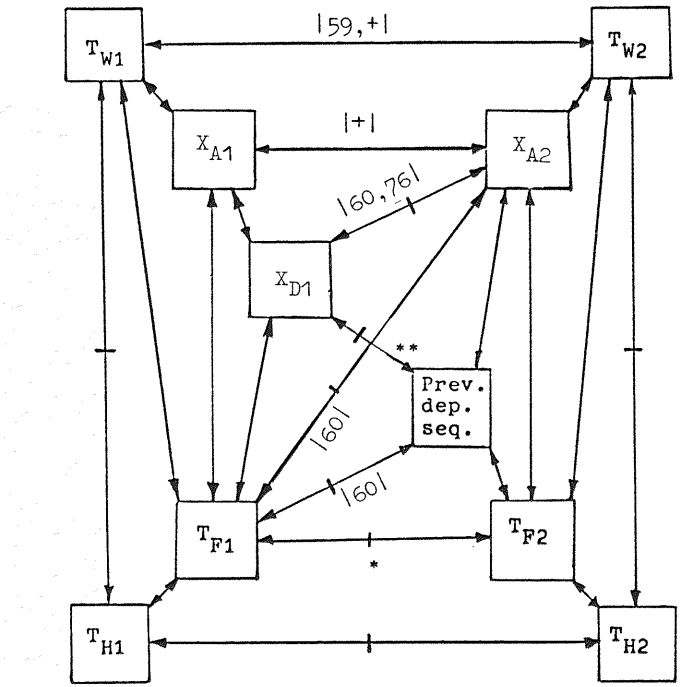


Bild 2.4 Beziehungen zwischen Schicksalsgrößen derselben Anforderung

Der überraschende Effekt der Unabhängigkeit der Durchlaufszeiten bei 2-stufigen Systemen wurde mit Hilfe des Konzepts der reversiblen Markoff-Prozesse von REICH |94-96| für Single--Server und von BURKE |60| für Multiserver-Systeme bewiesen und gilt auch (vgl BURKE |62|) für beliebige VF in Stufe 2.

BURKE zeigte in |59|,daß die Wartezeiten einer Anforderung in 2 aufeinanderfolgenden Single-Server Stufen abhängig sind.Mit Hilfe des Theorems von R.R.P. JACKSON und der virtuellen Wartezeit in Stufe 2 (Wartezeit einer Anforderung,die zum betrachteten Zeitpunkt ankommen würde) zeigte er durch explizite Rechnung,daß die Wartewahrscheinlichkeiten $W_i=P(T_{Wi}>0)$ in beiden

Stufen nicht unabhängig sind.Obwohl also der 2.Stufe ein reiner Zufallsverkehr angeboten wird,ist das spätere Schicksal einer Anforderung (Zahl X_{A2} der bei Ankunft in Stufe 2 dort angetroffenen Anforderungen und/oder Wartezeit T_{W2}) nicht unabhängig von ihrem früheren Schicksal in Stufe 1.
Im Output-Theorem von BURKE in |62| werden 2 Aussagen zusammengefaßt, nämlich daß

a) der Ausgangsprozeß einer Stufe M|M|n ein Poissonprozeß ist (aus |30|,vgl.2.2.1)
b) der Zustand der Stufe zu einer Zeit t unabhängig von ihrem vergangenen Ausgangsprozeß ist.

NELSON |88| leitete durch Faltung einen Ausdruck für die VF der Gesamtwartezeit in einem System mit allgemeinerer Struktur her, indem er Unabhängigkeit der einzelnen Wartezeiten annahm.
LEE |79|,der eine Durchlaufsgraphen-Technik entwickelte zur Bestimmung von Gesamtzeiten in einem Netz von Wartestufen, nimmt ebenfalls die Unabhängigkeit von Stufenzeiten an.

Bei seriellen Systemen mit mehr als 2 Stufen (und FIFO-Abfertigungsdisziplinen) wurde von REICH |95| die Unabhängigkeit der Durchlaufszeiten in den Einzelstufen für eine beliebige Anzahl von Single-Server Stufen bewiesen.Die Unmöglichkeit einer Verallgemeinerung auf beliebigstufige Multiserver-Systeme wurde von BURKE |61| gezeigt.Er bewies,daß im 3-stufigen Wartesystem

$$M \xrightarrow{\infty} M|1 \xrightarrow{\infty} M|n_2 \xrightarrow{\infty} M|1$$

zwar T_{F1} und T_{F2} unabhängig sind,ebenso T_{F2} und T_{F3},jedoch die Durchlaufszeit T_{F3} von T_{F1} abhängt!

Es seien nun die Beiträge hierüber von REICH und BURKE zusammengefaßt.

SATZ: Die Durchlaufszeiten T_{Fi} einer Anforderung in den m verschiedenen Stufen eines seriellen Markoff-Systems ohne Blockierung sind nur dann voneinander unabhängig, wenn alle Zwischenstufen (Stufe 2..m-1) Single-Server Stufen sind.

Dieser Satz wird etwas anschaulicher,wenn man bedenkt,daß in diesen Markoff'schen Systemen dann Abhängigkeiten bei den Durchlaufszeiten entstehen,wenn sich Anforderungen in einer Multiserver-Zwischenstufe durch unterschiedliche Bedienungszeiten gegenseitig "überholen" können.

Dann nämlich kann eine Anforderung,die sich z.B. durch eine große Bedienungszeit und damit auch große Durchlaufszeit in einer Stufe i-1 viele wartende Anforderungen als Konkurrenten in Stufe i geschaffen hat, von diesen in Stufe i überholt und dadurch in der Tendenz diese in verstärkter Zahl in Stufe i+1 antreffen ("Überholen nach vorherigem Rückstau").
Daß es sich hierbei nur um eine nachträgliche Deutung handeln kann,ist z.B. aus der Unabhängigkeit der Durchlaufszeiten bei 2-stufigen Systemen mit negativ-exponentiellen VF der Bedienungszeiten zu ersehen,die keineswegs von vornherein plausibel ist.Dagegen kann man jedoch sagen,daß beispielsweise bei konstanten Bedienungszeiten eine Unabhängigkeit der Durchlaufszeiten sicher ausgeschlossen werden kann.
Auf diesen Problemkreis,speziell bei 2-stufigen Anordnungen, wird in Kapitel 3 eingegangen werden.

2.2.3 Systeme mit Nicht-Markoff'schen Voraussetzungen

Wartesysteme in Serie ohne Blockierung bei Nicht-Markoff'schen Voraussetzungen wurden in der Literatur relativ selten betrachtet.Überdies wurden dabei meist Existenzbedingungen für stationäres Verhalten behandelt,was bei diesen Systemen ohne Blockierung zu für die Praxis meist trivialen Ergebnissen führt.
Hierzu zählen z.B. die Arbeiten von SACKS |99| über 2-stufige und von LOYNES |80| über beliebigstufige Single-Server Systeme

$$G \xrightarrow{\infty} G|1 \xrightarrow{\infty} G|1 \xrightarrow{\infty} \cdots \xrightarrow{\infty} G|1$$

mit beliebigen VF für den Ankunftsprozeß vor Stufe 1 und den Bedienungsprozeß in allen Stufen.
Durch Ansetzen der individuellen Zeiten von aufeinanderfolgenden Anforderungen in den verschiedenen Stufen (analog der Aufstellung der LINDLEY'schen Integralgleichung |40|) wurden Bedingungen hergeleitet,unter denen ein stationäres Verhalten des Systems gesichert und dieses eindeutig und unabhängig vom Ausgangszustand ist (Ergodizität).Danach ist (erwartungsgemäß) eine Stufe i für sich im statistischen Gleichgewicht (selbst wenn Teile des Gesamtsystems instationär werden),falls

$$\lambda_i \cdot h_i < 1$$

d.h. das Angebot an eine Single-Server Stufe i kleiner 1 ist.

MASTERSON und SHERMAN |84,85| untersuchten das Verhalten des obigen Single-Server Systems für Stufenzahlen $m \to \infty$ und zeigten daß der Erwartungswert der Abgangsabstände zwischen zwei aufeinanderfolgenden Anforderungen am Ausgang der letzten Stufe keinen endlichen Wert besitzt.
Neben diesen Fragen der Stationarität bei beliebigen VF wurden von einigen Autoren das 2-stufige Single-Server System

$$M \xrightarrow{\infty} VF_1|1 \xrightarrow{\infty} M|1$$

bei verschiedenen VF der Bedienungszeiten VF_1 in Stufe 1 betrachtet und insbesondere der Charakter der Wartezeit-VF in Stufe 2 untersucht.
GHOSAL |66| behandelte dieses System mit $VF_1=E_2$,also Erlang-2-verteilten Bedienungszeiten in Stufe 1,indem er Beziehungen aufstellte für individuelle Zeiten von Anforderungen in beiden Stufen inclusive der Ankunftsabstände und Freizeiten.Bei der Auswertung jedoch trifft er eine nichtgenannte implizite Annahme der Unabhängigkeit zwischen der zufälligen Durchlaufszeit einer Anforderung in Stufe 1 und deren Wartezeit in Stufe 2, was im allgemeinen nicht erfüllt ist und demzufolge keine exakte Lösung ergeben kann.
SUZUKI |102|,dessen 1.Stufe eine beliebige Bedienungszeit-VF besaß ($VF_1=G$),versuchte,die Wartezeit-VF in Stufe 2 mit Hilfe der Methode der eingebetteten Markoff-Kette zu bestimmen.Als Regenerationszeitpunkte für die Kette wählte er die Zeitpunkte zu denen eine Anforderung die 1.Stufe verläßt,so daß die Zeitintervalle zwischen zwei Betrachtungszeitpunkten mit den Ankunftsabständen in Stufe 2 identisch sind.Dann berechnete er die Übergangswahrscheinlichkeiten p_{ij} der Markoff-Kette und wendete KENDALL's Methode zur Lösung an.Jedoch muß der Autor offensichtlich übersehen haben,daß im allgemeinen Fall aufeinanderfolgende Ankunftsabstände vor Stufe 2 voneinander abhängig sind.Als (nicht exakte) Lösung für die Wartezeit-VF in Stufe 2 erhält SUZUKI eine negativ-exponentielle VF,was aufgrund seiner (falschen) Annahme der Unabhängigkeit sich ergeben muß,weil im System GI|M|1 die Wartezeit-VF unabhängig vom Eingangsprozeß negativ-exponentiell ist (vgl.|44|).
LOYNES |81| betrachtete genau wie GHOSAL das obige System mit $VF_1 = E_2$ und bewies -unter Einführung fiktiver Anforderungen

für Stufe 2 mit Bedienungszeit $T_{H2} \equiv 0$ und Anwendung spezieller Ergebnisse einstufiger Systeme- daß die Wartezeit-VF in Stufe 2 selbst keine negativ-exponentielle Funktion sein kann. Auch wies er auf den Widerspruch zwischen seinen Ergebnissen und denen von GHOSAL hin.

Serielle Wartesysteme mit nur konstanten Bedienungszeiten nehmen wegen des deterministischen Charakters ihrer Bedienungsprozesse eine Sonderstellung ein. FRIEDMAN [65] untersuchte ein beliebigstufiges System (ohne Blockierung) bei allgemeinem Eingangsprozeß für die 1. Stufe

$$G \xrightarrow{\infty} D|n_1 \xrightarrow{\infty} D|n_2 \xrightarrow{\infty} \cdots \xrightarrow{\infty} D|n_m$$

und bewies, daß die gesamte Durchlaufszeit (und damit die gesamte Wartezeit) einer Anforderung Nr.j im System unabhängig ist von der Reihenfolge der Stufen (FIFO-Abfertigungsdisziplin in jeder Stufe).
Er zeigte dies zunächst für zwei Stufen und erweiterte das Ergebnis sukzessiv auf beliebigstufige Systeme. Anhand ebenfalls 2-stufiger Systeme stellte er Dominanzbeziehungen auf zwischen zwei Stufen mit konstanten Bedienungszeiten h_1 bzw. h_2 und n_1 bzw. n_2 Bedienungseinheiten. Dabei dominiert Stufe 1 über Stufe 2 falls in der Reihenfolge 1-2 der Stufen bei beliebigem Eingangsprozeß in Stufe 1 nie eine Anforderung in Stufe 2 warten muß. Dies ist erfüllt, d.h. Stufe 1 dominiert über Stufe 2, wenn

$$\frac{h_2}{h_1} \leq \frac{n_2}{n_1} \Big| \text{ ganzzahlig abgerundet} \qquad (2.5)$$

Man kann zeigen, daß es 2-stufige Systeme gibt, bei denen weder Stufe 1 über Stufe 2 noch 2 über 1 dominiert. Dies bedeutet, daß dieses System nicht auf eine dominierende Stufe reduziert werden kann (es sei denn diese beiden Stufen sind Bestandteile eines Systems mit weiteren Stufen und jede der beiden Stufen werden von einer dritten Stufe dominiert). Jedoch bei identischer Zahl n von BE in allen Stufen ist stets eine Reduktion auf ein einstufiges System G|D|n möglich mit der konstanten Bedienungszeit

$$h = \max(h_1, h_2, .., h_m) \qquad (2.6)$$

Geht man von der FIFO-Abfertigungsdisziplin in jeder Stufe ab, so ist zwar nicht mehr die Ausgangsreihenfolge der Anforderungen gleich ihrer Eingangsreihenfolge, jedoch bleiben die Abgangsabstände und damit der Ausgangsprozeß unverändert.
Diese Ergebnisse von FRIEDMAN werden in Kapitel 4 verwendet, es sei hier nur noch auf die Untersuchungen entsprechender Systeme bei endlich großen Zwischenspeichern von AVI-ITZHAK, vgl. 2.3.2.2, verwiesen.

2.3 Systeme mit Blockierung

Bei Systemen mit Blockierung, wo aufgrund nur endlich großer Zwischenspeicher ein momentaner Rückstau über mehrere Stufen auftreten kann, ist es offensichtlich, daß eine große Korrelation zwischen den Belegungszuständen in den verschiedenen Stufen herrschen kann und es nicht mehr möglich ist, diese als Einzelstufen zu betrachten; deshalb sind auch die Zeiten von Anforderungen in aufeinanderfolgenden Stufen nicht unabhängig voneinander.
Durch Blindbelegung von BE in Stufe i sind im Mittel Y_{Bi} BE gleichzeitig blockiert (Blockierbelastung), es stehen also im Mittel nur $n_i - Y_{Bi}$ BE zur Verfügung, die pro Zeiteinheit entsprechend weniger Anforderungen abfertigen können.
Die wichtigste Kenngröße eines solchen Systems mit Blockierung ist deshalb die maximal verarbeitbare Rate λ_{max} (vgl. 1.3.3), die in vielen Veröffentlichungen Hauptziel der Untersuchungen ist.
Unter Markoff'schen Annahmen wurden hauptsächlich 2-stufige Systeme (auch Multiserver) betrachtet, bei mehr als 3 Stufen jedoch nur Single-Server Systeme mit gleichen Parametern für alle Stufen.
Mit allgemeinen Nicht-Markoff'schen Annahmen gibt es fast nur Single-Server Systeme, bei denen für mehr als 2 Stufen meist nur Stabilitätskriterien oder die Zahl der möglichen Systemzustände betrachtet wurden.
Wie auch bei Systemen ohne Blockierung wird im folgenden auf den Spezialfall rein Markoff'scher Verkehrsannahmen eingegangen. Dies erscheint hier durchaus sinnvoll, da aus dem allgemeinen Fall beliebiger VF nur sehr wenige neue Ergebnisse für den speziellen Fall negativ-exponentieller VF abgeleitet werden können.

2.3.1 Systeme mit rein Markoff'schen Voraussetzungen

Bei Systemen mit Blockierung ist es für diese Übersicht sinnvoll, den wichtigen Fall des maximalen Durchsatzes für sich zu betrachten.

2.3.1.1 Markoff'sche Systeme bei maximaler Durchsatzrate

Bei maximaler Durchsatzrate λ_{max} sind die BE der 1.Stufe immer voll belegt (Blockierung eingeschlossen, vgl.1.3.3). Dabei ist es unerheblich, wie dieser Betriebszustand erreicht wird. Er ist demzufolge unabhängig von einem evtl. vorhandenen Ankunftsprozeß und der Größe s_1 des Wartespeichers in Stufe 1; notwendig ist nur, daß bei Bedarf sofort eine neue Anforderung zur Verfügung steht (kein Ankunfts- sondern ein "Holprozeß"), was durch ein unendlich großes Reservoir beschrieben werden kann.
Neben der Wichtigkeit des maximalen Durchsatzes ist für seine bevorzugte Behandlung in der Literatur auch seine zum Teil leichtere Berechenbarkeit anzusehen (vom Ankunftsprozeß und von s_1 unabhängig). Bei der hierzu meist verwendeten Berechnungsmethode über ein Gleichungssystem für die stationären Zustandswahrscheinlichkeiten ist es immer möglich, das Problem auf ein einfacheres Gleichungssystem als bei $\lambda_Y < \lambda_{max}$ zurückzuführen.
Geht man direkt vom Grenzfalle maximalen Durchsatzes aus, so kann z.B. bei einem 2-stufigen System die Zustandsbeschreibung eindimensional erfolgen (BE in Stufe 1 bei Blockierung als "Verlängerung" des Zwischenspeichers) und man erhält (ohne Herleitung) anstatt

$$(n_1+s_1+1)\cdot(n_2+s_2+1) + n_1\cdot(s_1 + \frac{n_1+1}{2})$$

verschiedenen Zuständen im allgemeinen 2-stufigen System

$$M \xrightarrow{s_1} M|n_1 \xrightarrow{s_2} M|n_2$$

nur eine Zahl von

$$n_1 + n_2 + s_2 + 1 .$$

Dies sind selbst im Falle $n_1 = n_2 = 1$ je nach Größe des Zwischenspeichers nur 50-60% der erstgenannten Anzahl mit dem dort günstigsten Fall $s_1 = 0$.

Bei rein Markoff'schen Systemen (und solchen mit nur neg.-exponentiellen Teilphasen für Ankunftsabstände und Bedienungszeiten, z.B. E_k o.ä.) ist es prinzipiell immer möglich, das Gleichungssystem für die stationären Zustandswahrscheinlichkeiten anzusetzen (vgl.S.34, Bild 1.3). Ist keine explizite Lösung bekannt oder diese zu aufwendig, so kann das Gleichungssystem numerisch auf einem Rechner gelöst werden, solange die Zahl der Unbekannten den Arbeitsspeicher des Rechners nicht sprengt. Darüber hinaus ist die Ermittlung eines numerisch rekursiven Verfahrens -sofern möglich- von großem Vorteil.
Aus den Zustandswahrscheinlichkeiten lassen sich dann in bekannter Weise die Belastungen Y_i der Stufen gewinnen (Blockierung nicht enthalten), woraus sich aus Stationaritätsgründen direkt der maximale Durchsatz

$$\lambda_{max} = Y_i \cdot \varepsilon_i \qquad i = 1..m \quad (2.7)$$

ergibt, wobei $\varepsilon_i = 1/h_i$.

Neben dieser wichtigen Methode der λ_{max}-Bestimmung gibt es eine weitere, bei der die mittlere Gesamtbelegungszeit der Anforderungen in der BE der Stufe 1 bestimmt wird (vgl.|82|), die im Prinzip für beliebige VF gilt: da jede BE der 1.Stufe immer belegt ist, gilt für die Durchsatzrate der 1.Stufe und damit des Systems

$$\lambda_{max} = n_1 \cdot \frac{1}{E(T_{H1}) + E(T_{B1})} \qquad (2.8)$$

wobei

$$E(T_{B1}) = p_{B1} \cdot t_{B1} \qquad \text{bei } \lambda_Y = \lambda_{max} \qquad (2.9)$$

Zum Beispiel bei 2-stufigen Systemen gilt bei $n_1 = 1$ und negativ-exponentiell verteilten Bedienungszeiten in Stufe 2

$$t_{B1} = \frac{h_2}{n_2} ,$$

womit das Problem auf die Bestimmung der Blockierwahrscheinlichkeit zurückgeführt werden kann.
Da die maximale Durchsatzrate einen Grenzwert darstellt, kann diese natürlich auch aus Ergebnissen gewonnen werden, die für allgemeinen Durchsatz gelten, nämlich

- bei $s_1 \to \infty$ als maximal zulässige Rate für Stationarität
- bei $0 \leq s_1 < \infty$ als verarbeitete Rate λ_Y bei $\lambda \longrightarrow \infty$

(vgl. 2.3.1.2).

Im folgenden wird nun auf die zugehörige Literatur eingegangen und zwar auf
- beliebige 2-stufige Systeme
- 3-stufige Single-Server Systeme
- beliebigstufige Single-Server Systeme.

● Beliebige 2-stufige Systeme unter Markoff'schen Voraussetzungen werden bei maximaler Durchsatzrate in dieser Arbeit mit

$$\rightarrow M|n_1 \xrightarrow{s_2} M|n_2$$

symbolisch bezeichnet (Symbol f.Reservoir vgl.S.42).

Als erste Veröffentlichung über serielle Wartesysteme mit Blockierung zählt die von HUNT (1956) |73|,der u.a. für obiges System mit $n_1 = n_2 = 1$ den maximalen Durchsatz bestimmte.

$$\lambda_{max} = \frac{\varepsilon_1^{s_2+2} - \varepsilon_2^{s_2+2}}{\varepsilon_1^{s_2+3} - \varepsilon_2^{s_2+3}} \cdot \varepsilon_1 \cdot \varepsilon_2 \qquad (2.10a)$$

mit den beiden Sonderfällen

$$\lambda_{max} = \frac{\varepsilon_1+\varepsilon_2}{\varepsilon_1^2+\varepsilon_1\varepsilon_2+\varepsilon_2^2} \cdot \varepsilon_1 \cdot \varepsilon_2 \quad \text{für } s_2 = 0 \qquad (2.10b)$$

$$\lambda_{max} = \frac{s_2+2}{s_2+3} \cdot \varepsilon \quad \text{für } \varepsilon_1 = \varepsilon_2 = \varepsilon \qquad (2.10c)$$

Dieses Ergebnis wird auch in Kapitel 5 benötigt,wo ein Approximationsverfahren für das gleiche System,jedoch mit beliebigen VF der Bedienungszeiten entwickelt wird.

Dabei ging HUNT von den Zustandsgleichungen für das entsprechende System bei $s_1 \rightarrow \infty$ aus und bestimmte daraus die Stabilitätsgrenze für stationäres Verhalten,die z.B. bei 2 seriell angeordneten gleichen BE ohne Zwischenspeicher bei $Y_1 = 2/3$ erreicht wird.
Obige Formel ist in ε_1 und ε_2 symmetrisch,was gleichen maximalen Durchsatz bei Vertauschen der BE bedeutet.
MAKINO |82| setzte direkt die Zustandsgleichungen für obiges System mit beliebigem n_1 und n_2 an und leitete daraus eine -natürlich wesentlich kompliziertere aber noch explizite- Formel für den maximalen Durchsatz ab.

● Bei 3-stufigen Systemen wurden nur Single-Server betrachtet.
HUNT |73| bestimmte,wiederum über die Zustandswahrscheinlichkeiten,für $s_2 = s_3 = 0$ explizit den maximalen Durchsatz,bei dem die Ausnutzung der 1.Stufe $Y_1 = 22/39 \approx 0.564$ beträgt.

MAKINO |82| kommt auf ähnliche Weise zu einer expliziten Formel, die in ε_1 und ε_3 symmetrisch ist.Nach einigen Umformungen kann man zeigen,daß seine allgemeine Formel mit dem Ergebnis von HUNT übereinstimmt.
Als Ergebnis aus |82| sei noch bemerkt,daß bei gegebenem Tripel $\{\varepsilon_a,\varepsilon_b,\varepsilon_c\}$ diejenige Reihenfolge der BE den größten maximalen Durchsatz ergibt, bei der der mittleren BE die höchste Servicerate zugeteilt wird:

$$\varepsilon_2 = \max(\varepsilon_a,\varepsilon_b,\varepsilon_c) \qquad (2.11)$$

Dieses Ergebnis kann dahingehend interpretiert werden,daß bei einer solchen Reihenfolge die BE in Stufe 2 bevorzugt als ein Zwischenpuffer zwischen Stufe 1 und 3 dienen kann.

HILLIER und BOLING |71| behandelten den Aspekt der Reihenfolge ausführlicher auf der Grundlage von HUNT's Ergebnissen und weiterer exakter Rechnungen bis 4 Stufen und mit Zwischenspeichern.

HATCHER |67| setzte in bekannter Weise für das allgemeine 3-stufige Single-Server System mit Zwischenspeichern die Gleichgewichtsbedingungen für die $(s_2+3)\cdot(s_3+3)-1$ zweidimensionalen Zustände an und leitete daraus explizite aber doch schon relativ unhandliche Formeln ab.

● HILLIER und BOLING |72| entwickelten für beliebigstufige Single-Server Systeme mit negativ-exponentiellen Bedienungszeiten in allen Stufen ein elegantes und effizientes Näherungsverfahren zur numerisch-iterativen Bestimmung der maximalen Durchsatzrate und der Gesamtzahl der Anforderung im System.

2.3.1.2 Markoff'sche Systeme bei allgemeiner Durchsatzrate

Bei dieser Betriebsweise haben einzelne BE der 1.Stufe "Freizeiten"; es ist z.B.möglich, daß Stufe 1 ganz leer ist.Deshalb müssen bei der Berechnung mehr Zustände berücksichtigt werden als bei maximalem Durchsatz.

Es bietet sich in diesem Falle an, die in der Literatur behandelten Systeme nach der Größe des Wartespeichers in Stufe 1 zu gliedern:

- Systeme ohne Wartespeicher in Stufe 1 ($s_1=0$)
- Systeme mit $s_1 \rightarrow \infty$
- Systeme mit $0 \leq s_1 < \infty$

● Zu den am meisten untersuchten Systemen gehören solche mit $s_1=0$, wo also Stufe 1 (abgesehen von der Blockierung, die ja auch Warten bedeutet) als reines Verlustsystem arbeitet. In diesem Falle ist bei festem Restsystem die Gesamtzahl der Zustände noch am kleinsten, was sich natürlich günstig auf die Lösbarkeit auswirkt. Es sei nun das System

$$M \xrightarrow{0} M|1 \xrightarrow{s_2} M|1$$

mit beliebig großem Zwischenpuffer der Größe s_2 betrachtet, das $2s_2+5$ verschiedene Zustände besitzt.
MORSE |86| berechnete für die einfachen Fälle $s_2 = 0,1$ die Zustandswahrscheinlichkeiten und daraus weitere Größen wie die mittlere Gesamtzahl von Anforderungen im System.
Eine Erweiterung dieser Ergebnisse auf $s_2=2$ und ein Lösungsschema für größere Zwischenspeicher wurde von STANGE |101| angegeben, der jedoch für $s_2=7$ wegen des geringeren Aufwands eine numerische Lösung des Gleichungssystems vorzieht. Dabei wird gezeigt, daß bei dieser Speichergröße (und negativ-exponentiell verteilten Bedienungszeiten) für Angebot $A<0.8$ nur noch eine geringe Blockierung herrscht. Deshalb kann man oft in erster Näherung mit einem unendlich großen Zwischenspeicher rechnen. Für dieses System ohne Blockierung gibt STANGE die Zustandswahrscheinlichkeiten und Folgegrößen an.

Bei WDOWIN |106| (in Russisch) ist in Erweiterung zu obigem System in der 2. Stufe eine beliebige Zahl von BE zugelassen. Ausgangspunkt sind ebenfalls die Gleichungen für die stationären Zustandswahrscheinlichkeiten, die mit Hilfe von Matrizen ausgewertet werden. Dabei findet der Fall $n_2=1$ naturgemäß eine besondere Beachtung.
Das transiente Verhalten des obigen Systems mit $s_2=0$ untersuchte PATTERSON |92| durch Aufstellen des Differentialgleichungssystems für die zeitabhängigen Zustandswahrscheinlichkeiten. Er zeigte für den Fall des bei t=0 leeren Systems, daß ihre Lösung neben dem Term für $t \rightarrow \infty$ gedämpfte Sinus- und Cosinus-Anteile enthält.

● Im Falle $s_1 \rightarrow \infty$ liegt ein reines Wartesystem vor. Als Nachteil kann hier die unbegrenzte Zahl der Zustandswahrscheinlichkeiten gelten, dem jedoch der Wegfall von Unsymmetrien an gewissen Rändern des Zustandsraums gegenübersteht, was eine formelmäßige Lösbarkeit oft begünstigt.
MORSE |86| berechnete die Zustandswahrscheinlichkeiten im System

$$M \xrightarrow{\infty} M|1 \xrightarrow{0} M|1 \ .$$

Außerdem bestimmte er die Gesamtzahl E(X) von Anforderungen zu

$$E(X) = \frac{4A(2-A^2)}{(2+A)\cdot(2-3A)} \qquad \text{mit } A = \lambda \cdot h, \ h_1=h_2=h \qquad (2.12)$$

Hieraus ist sofort die Stationaritätsgrenze A=2/3 zu ersehen, die wegen $s_1 \rightarrow \infty$ existiert und dem maximalen Durchsatz entspricht.
Die Richtigkeit der angeführten Ergebnisse für die Zustandswahrscheinlichkeiten muß jedoch (aufgrund von RÖCK |97|, vgl. unten) bezweifelt werden, wie schon von STANGE |100| bemerkt wurde.
MAKINO |83| behandelte obiges System mit beliebiger Größe des Zwischenspeichers approximativ und bestimmte - ähnlich wie er es für die einstufigen Systeme $M|G|1$, $E_2|M|1$ und $E_2|E_2|1$ vollzog - die erzeugende Funktion für die Momente der VF des Ausgangsprozesses aus diesem 2-stufigen System. Zur approximativen Berechnung der 2-stufigen Anordnung bestimmt MAKINO nun ein einstufiges $M|G|1$-System mit gleicher momenterzeugender Funktion des Ausgangsprozesses, was auf eine vom Angebot abhängige VF der Bedienungszeiten führt.
Bezüglich des Ausgangsprozesses einer 3-stufigen Anordnung ohne Zwischenspeicher zeigte er, daß der Varianzkoeffizient kleiner ist als beim entsprechenden System mit 2 Stufen.
● Über Systeme mit endlich großem Wartespeicher in Stufe 1 ist kein Beitrag bekannt. Jedoch wurde von RÖCK |97| parallel zu vorliegender Arbeit ein Rechenprogramm erstellt zur numerischen Berechnung des allgemeinen 2-stufigen Markoff-Systems

$$M \xrightarrow{s_1} M|n_1 \xrightarrow{s_2} M|n_2 \qquad \text{mit } s_i \geq 0, \ n_i \geq 1 \quad i=1,2 \ .$$

Viele in der Literatur behandelten Systeme stellen hiervon Spezialfälle dar.

2.3.2 Systeme mit Nicht-Markoff'schen Voraussetzungen

In diesem Abschnitt sollen Systeme mit Blockierung bei beliebigen Annahmen bezüglich der VF der Bedienungszeiten wie auch der Ankunftsabstände behandelt werden. Die in 2.2.2 besprochenen Systeme mit negativ-exponentiellen VF sind hiervon Spezialfälle. Es ergeben sich jedoch nachstehend für diese keine neuen Ergebnisse.

Es wird zunächst der Betrieb bei maximalem Durchsatz betrachtet und dann der Fall allgemeiner Durchsatzraten, wo -abgesehen vom Grenzfalle nur konstanter VF- in der Literatur praktisch nur 2-stufige Single-Server Systeme behandelt wurden.

2.3.2.1 Nicht-Markoff'sche Systeme bei maximaler Durchsatzrate

Der Durchsatz durch ein System mit Blockierung in allen Stufen erreicht dann seinen maximalen Wert, wenn alle BE der 1.Stufe immer voll belegt sind (vgl.1.3.3).
In 2.3.1.1 wurden Methoden und Ergebnisse geschildert, die hierzu für negativ-exponentielle VF bekannt sind. Auf beliebige VF anwendbar ist das dort angedeutete Verfahren zur Bestimmung des maximalen Durchsatzes über die mittlere Aufenthaltszeit einer Anforderung in einer BE der Stufe 1. Diese wurde von MAKINO |82| für 2 seriell angeordnete BE (vgl.Bild 2.5) angegeben.

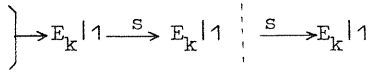
Bild 2.5 a) Strukturelle b) Symbolische Darstellung von 2 Servern in Serie bei maximaler Durchsatzrate

Da in diesem System dann eine Anforderung BE_1 belegt, wenn ihre Vorgängerin (nach einer eventuellen Blockierphase) ihrerseits den Service in Stufe 2 beginnt, ist die Gesamtbelegungszeit (Bedienung und Blockierung) einer Anforderung auf BE_1 gleich dem Maximalwert aus beiden zufälligen Bedienungszeiten. Damit ergibt sich für den Erwartungswert der Gesamtbelegungszeit der BE_1

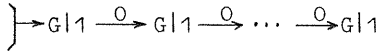
$$E(T_{H1}+T_{B1}) = E(\max\{T_{H1},T_{H2}\}) \qquad (2.13)$$

Dieser kann aus den beiden VF für die Bedienungszeiten bestimmt werden.

Speziell für negativ-exponentielle VF erhält MAKINO natürlich die gleichen Ergebnisse wie HUNT, während er explizit zeigt, daß bei weniger stark schwankenden Bedienungszeiten (E_2) der maximale Durchsatz größer wird.
HILLIER und BOLING |72| bestimmten numerisch für das 2 bzw.3--stufige System



den maximalen Durchsatz durch Zurückführung auf rein Markoff'sche Voraussetzungen mittels einer detaillierteren Zustandsbeschreibung. Numerisch exakte Werte wurden angegeben für k=1..10 und s=0..10.
HILDEBRAND |69,70| untersuchte das System ohne Zwischenspeicher



und behandelte Stationaritäts- und Ergodizitätsfragen bei beliebigem Durchsatz und $s_1 \to \infty$, aber auch Fragen des maximalen Durchsatzes, den er für 3 und 4 Stufen im Falle negativ-exponentieller VF explizit angibt.
MUTH |87| gibt für den maximalen Durchsatz im gleichen System eine einfache und deshalb auch oft grobe Abschätzungsformel an:

$$\frac{1}{E(\max T_{H1},T_{H2},..)} \leq \lambda_{max} \leq \frac{1}{\max(h_1,h_2,...)} \qquad (2.14)$$

Dabei ist die obere Grenze der Durchsatz des Systems bei unendlich großen Zwischenspeichern, wo sich Schwankungen der Bedienungszeiten nicht mehr auf λ_{max} auswirken, sondern nur noch die Mittelwerte der Bedienungszeiten ; die untere Grenze ist ein Fall, der bereits von HUNT behandelt wurde. Dort wird angenommen, daß die Anforderungen, bevor sie in die nächste Stufe gelangen, erst warten müssen, bis alle Anforderungen im System in der jeweiligen Stufe ihren Service beendet haben, um dann gemeinsam in die jeweils folgende Stufe zu gelangen (System mit gemeinsamen zustandsabhängigem Takt). Bei konstanten Bedienungszeiten fallen beide Grenzwerte zusammen.
MUTH gibt bis zu 10 Stufen mit gleichem Mittelwert der Bedienungszeiten die untere Grenze an und zwar für verschiedene hypoexponentielle VF inclusive E_k.

Für die Bestimmung der Zahl der Zustände im beliebigstufigen System

$$\}\rightarrow G|1 \xrightarrow{s_2} G|1 \xrightarrow{s_3} \cdots \xrightarrow{s_m} G|1$$

entwickelte HAYDON |68| ein rekursives Verfahren,während PATTERSON |92| für dieses System auch bei beliebigem Durchsatz und Speichergröße s_1 schon früher den Weg einschlug,diese Zahl mit Hilfe von linearen simultanen Differenzengleichungen zu bestimmen.
In SWOBODA und ROSENBOHM |105| wurde eine serielle Anordnung beliebig vieler BE untersucht,bei der die Anforderungen zusätzlich mit unabhängigen Wahrscheinlichkeiten hinter jeder BE das System verlassen können.Dieses Verkehrsmodell wurde für einen Prozessor entwickelt,bei dem die Bearbeitung der Befehle durch autonome Unterwerke erfolgt,deshalb wurde für die Bedienungszeiten eine spezielle Mehrpunkt-VF angenommen.Für dieses System wurde ein effektives und recht genaues Näherungsverfahren zur Bestimmung der maximal verarbeitbaren Rate angegeben.

2.3.2.2 Nicht-Markoff'sche Systeme bei allgemeiner Durchsatzrate

Hier wurden vornehmlich 2-stufige Single-Server Systeme ohne Zwischenspeicher betrachtet:

$$M \xrightarrow{\infty} G|1 \xrightarrow{0} G|1$$

SUZUKI |103| berechnete dieses System mit einer eingebetteten Markoff-Kette und leitete daraus neben Ergodizitätsbedingungen die erzeugende Funktion und den Mittelwert der Warteschlangenlänge in Stufe 1 ab, ebenfalls die Blockierwahrscheinlichkeit von Anforderungen.Durch Ansetzen der individuellen Zeiten von Anforderungen leitete er in |104| auch für beliebige VF der Ankunftsabstände vor Stufe 1 Bedingungen für Stationarität und Ergodizität ab.
Obiges System wurde ausführlich auch von AVI-ITZHAK und YADIN |56| behandelt.Sie betrachteten Test-Anforderungen und verwendeten vorhandene Ergebnisse für ein spezielles M|G|1-System bei dem jeweils die erste Anforderung einer Arbeitsperiode eine beliebige andere Bedienungszeit-VF besitzt (System mit "special service").Es wurden u.a. die Wartezeiten in Stufe 1 und die Durchlaufszeiten in Stufe 1 und im Gesamtsystem

bestimmt (Mittelwert,momenterzeugende Funktion).Es wurde gezeigt daß für die beiden Fälle M-M und D-D die mittlere Durchlaufszeit durch das System (und damit die Gesamtzahl der Anforderungen) unabhängig von der Reihenfolge der beiden BE ist.
Das transiente Verhalten dieses Systems wurde von PRABHU |93| mit Hilfe der Theorie der Erneuerungsprozesse betrachtet (vgl. z.B. FELLER |9|,während CHANG |63| in obigem System einen E_k-Eingangsprozeß annahm und verschiedene Verkehrsgrößen mit einer eingebetteten Markoff-Kette bestimmte.
Systeme mit endlich großem Zwischenspeicher

$$M \xrightarrow{\infty} G|1 \xrightarrow{s_2} M|1$$

wurden von NEUTS |89,90| betrachtet,wobei jedoch aus Berechnungsgründen eine negativ-exponentielle VF in Stufe 2 angenommen wurde.Der mathematische Lösungsaufwand ist immens und die Ergebnisse größtenteils so kompliziert,daß eine sinnvolle Anwendung in der Praxis mit größerer Sicherheit ausgeschlossen werden kann (vgl. Bemerkung von BURKE hierzu in 2.1).
Als besonders erwähnenswert erscheint bei diesem System die mögliche exakte Bestimmung der Blockierwahrscheinlichkeit und damit wegen $t_{B1} = h_2$ der maximalen Durchsatzrate,wozu NEUTS ein rekursives Lösungsverfahren angibt mit s_2+1 aus der VF G in Stufe 1 zu berechnenden Koeffizienten.
Deshalb wurde für dieses System,das bei allgemeiner Durchsatzrate und beliebiger hypoexponentieller VF in Stufe 2 im Rahmen vorliegender Arbeit behandelt wird,ein Näherungsverfahren entwickelt (vgl. Kapitel 6).
Die Sonderstellung serieller Wartesysteme mit nur konstanten VF der Bedienungszeiten gilt nicht nur für Systeme ohne Blockierung (vgl. 2.2.3),sondern auch für den hier betrachteten Fall endlich großer bzw. ohne Zwischenspeicher.
AVI-ITZHAK |57| zeigte -unabhängig von FRIEDMAN- daß im System

$$G \rightarrow D|n \xrightarrow{s_2} D|n \xrightarrow{s_3} D|n \rightarrow \cdots \xrightarrow{s_m} D|n$$

mit gleicher Zahl von BE in allen Stufen

- die gesamte Durchlaufszeit einer Anforderung unabhängig ist von der Reihenfolge der BE und den Zwischenspeichergrößen
- ein 1-stufiges Ersatzsystem mit gleicher Verzögerungszeit in einfacher Weise (analog wie bei FRIEDMAN) angegeben werden kann.

Dies zeigte er auch für den Fall ungleicher Zahl von BE für das System

$$G \xrightarrow{\infty} D|n_1 \xrightarrow{0} D|n_2$$

und wies auf die naheliegende Möglichkeit hin,daß obige Aussagen auch für beliebigstufige Systeme mit verschiedenen n_i gilt, was er aber nicht beweisen konnte.
Hierzu kann gesagt werden,daß es mit dem in dieser Arbeit verwendeten Simulationsprogramm nicht gelang,diese Vermutung zu widerlegen.

2.4 Verwandte Systeme

Zur Abrundung wird noch ein kurzer Ausblick gegeben auf einige Systeme,die den hier behandelten seriellen Anordnungen in der Struktur,den Berechnungsmethoden und vor allem in den Ergebnissen verwandt sind.

2.4.1 Geschlossene zyklische Wartesysteme

Geschlossene Systeme können dadurch gekennzeichnet werden,daß immer eine konstante Zahl von Anforderungen in ihnen enthalten ist.Geschlossene zyklische Systeme entstehen z.B. dadurch,daß rein seriell angeordnete Wartestufen zu einem Ring geschlossen werden in dem dann eine bestimmte Zahl von Anforderungen zyklisch umläuft.Diese Systeme treten z.B. bei Rechnermodellen für Multiprogramming-Betrieb auf (vgl.z.B. |50|),einen Einblick in bestehende Ergebnisse findet man in COFFMAN |48|.
Über Äquivalenzbeziehungen zu speziellen offenen Systemen wird in Abschnitt 5.3 berichtet.

2.4.2 Serielle Wartesysteme mit Phasen überlappter Belegung (Übergabezeiten)

Bei allen betrachteten Modellen wurde angenommen,daß eine Anforderung gleichzeitig nur eine Einheit (Warteeinheit,BE) belegt.Erfolgt jedoch eine gleichzeitige Belegung solcher Einheiten zum Zwecke der Informationsübergabe oder Zusammenarbeit, so muß bei endlicher Zahl dieser Einheiten diese überlappte Belegung mitberücksichtigt werden.Solche komplexeren Modelle wurden parallel zu vorliegender Arbeit behandelt |77,78|.

2.4.3 Netze aus Wartestufen

Der allgemeine Fall der Anordnung von mehreren Einzelstufen ist der eines Netzes,bei dem die Wege des Durchlaufs von Anforderungen in jedem "Knoten" beginnen und enden können und z.B. durch Verzweigungswahrscheinlichkeiten gekennzeichnet sind. J.R.JACKSON |74| zeigte über die Zustandsgleichungen,daß die von R.R.P. JACKSON für serielle Wartesysteme ohne Blockierung bei rein Markoff'schen Voraussetzungen bewiesene Unabhängigkeit der Zustandswahrscheinlichkeiten der Einzelstufen auch bei beliebigen derartigen Netzen gilt.
Ist die Anordnung der einzelnen Wartestufen so,daß die aus verschiedenen Richtungen eintreffenden Anforderungen eines Knotens für sich in unabhängigen Poissonprozessen ankommen (d.h. z.B. ohne Rückkopplungen),so kann die Berechnung der Einzelstufen (Knoten) mit Hilfe der Poissoneigenschaft des Ausgangsprozesses einer Stufe M|M|n erfolgen.Dabei gilt die Poissoneigenschaft auch separat für jede Verzweigungsrichtung,falls diese unabhängig ausgewählt werden.Berücksichtigt man nun den Satz über die Unabhängigkeit des Zustands einer M|M|n-Stufe von ihrem vergangenen Ausgangsprozeß (REICH |94|,BURKE |62|), so kann man auch für diese speziellere Strukturen die Zustandswahrscheinlichkeit des gesamten Netzes als Produkt der Einzelwahrscheinlichkeiten darstellen.
Die bei allgemeinen Netzstrukturen auftauchenden Fragen bezüglich der Ausgangsprozesse behandelte BURKE |62|,der auch die erstaunliche Tatsache diskutierte,daß die Zustandswahrscheinlichkeiten der Einzelstufen zwar so sind,als ob ihr Eingangsprozeß ein Poissonprozeß wäre,obwohl dieser aber im allgemeinen (bei Rückkopplungen etc) kein Poissonprozeß ist!
Für die Berechnung von allgemeinen Netzen müssen häufig Unabhängigkeitsannahmen getroffen werden,die sich entweder auf die Bedienungszeiten in aufeinanderfolgenden Stufen beziehen (z.B. Übertragungszeiten einer Nachricht) oder auf die Durchlaufszeiten in aufeinanderfolgenden Abschnitten des Netzes, vgl. z.B. KLEINROCK |12|,HERZOG |11|.

3 VERFOLGUNG VON ANFORDERUNGEN DURCH DAS GESAMTSYSTEM

In Kapitel 1 und 2 wurde eine Einführung in Systeme mit seriellem Warten gegeben sowie eine Übersicht über die in der Literatur behandelten Systeme.
In diesem Kapitel 3 werden nunmehr 2 seriell angeordnete Wartestufen mit unendlich großen Wartespeichern betrachtet. Dabei können die Anforderungen zusätzlich bereits nach Durchlaufen der 1. Stufe dieses System mit vorgebbarer Wahrscheinlichkeit verlassen:

$$M \xrightarrow{\infty} M|1 \;\downarrow \xrightarrow{\infty} M|n_2$$

Die Verzweigungsmöglichkeit ist - in Erweiterung zu der Kennzeichnung von mehrstufigen Systemen (vgl. Abkürzungsverzeichnis) - durch einen zusätzlichen abgehenden Pfeil gekennzeichnet: die 1. Stufe ist eine Single-Server Stufe mit negativ-exponentiell verteilten Ankunftsabständen, während Stufe 2 eine beliebige Zahl von Bedienungseinheiten (BE) besitzen darf. Die Bedienungszeiten in beiden Stufen seien negativ-exponentiell verteilt, jedoch gelten einige der Ergebnisse auch für beliebige VF in Stufe 2.
Durch Verfolgen von Test-Anforderungen während ihres Durchlaufs durch das ganze System wird das Schicksal einer Anforderung in der 2. Stufe bestimmt in Abhängigkeit aller ihrer möglichen Teilschicksale in Stufe 1 (Abschnitt 3.2). Es wird gezeigt, daß das Schicksal einer Anforderung in Stufe 2 unabhängig von dem konkreten Wert (>0) ihrer Wartezeit in Stufe 1 ist, also nur davon abhängt, ob die Anforderung in Stufe 1 gewartet hat oder nicht. Deshalb kann in 3.3 die VF der gesamten Wartezeit von Anforderungen in Stufe 1 und/oder 2 durch Faltung bedingter Wartezeit-VF bestimmt werden.
Schließlich werden in Abschnitt 3.4 auch Anforderungen untersucht, die eine bestimmte bekannte Anzahl von Anforderungen bei ihrer Ankunft in der ersten Stufe angetroffen haben. Indem alle möglichen Wege in einem RANDOM WALK-Diagramm betrachtet werden, wird die Zahl der bei Ankunft in Stufe 2 dort angetroffenen Anforderungen bestimmt, wie auch die daraus sich ergebenden Warte- und Durchlaufszeiten.

3.1 Einführung

Bei mehrstufigen Wartesystemen gibt es außer den bei einstufigen Systemen auftretenden Verkehrsgrößen (vgl. 1.3.2) auch Größen, die sich auf das gesamte System beziehen, sowie insbesondere Größen bezüglich des Durchlaufs von Anforderungen durch das Gesamtsystem (vgl. 1.3.2). Dazu gehört z.B. die VF der gesamten Durchlaufszeit, wie auch die VF der gesamten Verzögerungszeit, die bei Systemen ohne Blockierung der gesamten Wartezeit entspricht, in der auch die Wahrscheinlichkeit enthalten ist, daß eine Anforderung irgendwo im System warten muß.
In Abschnitt 2.2.2 wurde eine ausführliche Übersicht über die in der Literatur vorhandenen diesbezüglichen Ergebnisse gegeben, sie kann deshalb in diesem Kapitel als bekannt vorausgesetzt werden und dient als eigentliche Einführung in den hier behandelten Problemkreis.

3.1.1 Beschreibung des Systems

Das hier behandelte System besteht aus 2 seriell angeordneten Wartestufen mit unbegrenzt großen Speichern, bei dem zusätzlich die Anforderungen mit einer bestimmten Wahrscheinlichkeit $1-\eta$ dieses System bereits nach Stufe 1 verlassen können (vgl. Bild 3.1).

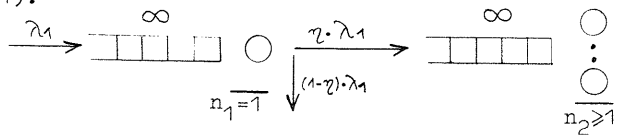


Bild 3.1 Das behandelte System

Die erste Stufe besitzt $n_1=1$ BE, während Stufe 2 eine Multiserver-Stufe sein darf. Dieses System ist u.a. ein Teilmodell, mit Hilfe dessen ein 2-stufiges Wartesystem ohne Blockierung, aber mit beliebig vielen parallelen "2.Stufen" (Richtungen) vollständig berechnet werden kann, wenn $\eta\cdot\lambda_1$ derjenige Verkehrsanteil ist, der in die gerade betrachtete Richtung fließt und $(1-\eta)\cdot\lambda_1$ die Summe der in andere, momentan nicht betrachteten Richtungen fließenden Verkehre. Dieses Modell entspricht einer Datenvermittlung mit einer zentralen Informationsverarbeitung und verschiedenen abgehenden Leitungsbündeln (vgl. 1.4.2).
Es wird angenommen, daß die Anforderungen in der 1. Stufe in einem Poissonprozeß mit der Ankunftsrate λ_1 eintreffen und später mit der unabhängigen Wahrscheinlichkeit η die betrachtete

Richtung einschlagen, so daß in Stufe 2 eine Ankunftsrate

$$\lambda_2 = \eta \cdot \lambda_1 \qquad (3.1)$$

herrscht. Die Bedienungszeiten T_{Hi} einer Anforderung (i=1,2) der betrachteten Richtung seien in beiden Stufen voneinander unabhängig und negativ-exponentiell verteilt mit den VF

$$H_i(\leq t) = 1 - e^{-\varepsilon_i t} \qquad (3.2)$$

und den Mittelwerten

$$E(T_{Hi}) = \frac{1}{\varepsilon_i} = h_i \qquad (3.3)$$

Die angebotenen Verkehre A_i (Angebote) sind

$$A_i = \lambda_i \cdot h_i \qquad (3.4)$$

und die Ausnutzungen

$$\varrho_i = \frac{A_i}{n_i} = \frac{\lambda_i}{\mu_i} \quad \text{mit } \mu_i = n_i \cdot \varepsilon_i \qquad (3.5)$$

Beide Stufen sollen sich in statistischem Gleichgewicht befinden, so daß

$$\lambda_1 < \min(\mu_1, \frac{\mu_2}{\eta}) \qquad (3.6)$$

Betrachtet man eine beliebige Anforderung, so erhält man das folgende Zeitdiagramm:

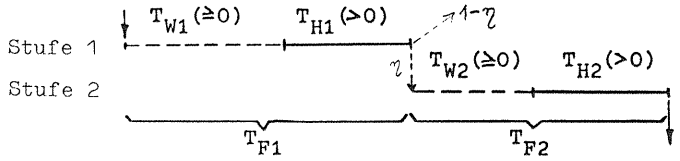


Bild 3.2 Allgemeines Zeitdiagramm für einen Durchlauf

T_{Wi} ist die zufällige Wartezeit, T_{Hi} die Bedienungszeit und T_{Fi} die Durchlaufszeit in Stufe i. Die gesamte Wartezeit einer Anforderung beträgt

$$T_W = T_{W1} + T_{W2} \qquad (3.7)$$

und die Gesamt-Wartewahrscheinlichkeit

$$W \overset{def}{=} P(T_W > 0) \qquad (3.8)$$

Die Abfertigungsdisziplinen in beiden Stufen sind beliebig, solange keine VF der Warte- und Durchlaufszeiten betrachtet werden.

Soweit diese VF betrachtet werden, wird Abfertigung in Ankunftsreihenfolge (FIFO) vorausgesetzt.
Aus der Warteschlangentheorie ist nun bekannt (vgl. 2.2.2.1 und 2.4.3), daß die Zustandswahrscheinlichkeiten in Stufe 2 von denen in Stufe 1 zu gleichen Zeitpunkten voneinander unabhängig sind und zwar so, als ob in Stufe 2 ein Poissonverkehr ankäme (Berechnung nach M|M|n, Theorem von R.R.P. bzw. J.R. JACKSON). Außerdem folgt aus dem Theorem von BURKE, daß der Ausgangsprozeß aus Stufe 1 ein Poissonprozeß ist und damit -bei Zuhilfenahme eines Satzes über Verzweigungen von Poissonprozessen- der Ankunftsprozeß in Stufe 2 auch tatsächlich ein Poissonprozeß ist. Diese bekannten Aussagen beziehen sich auf das System, während in diesem Kapitel neue Aussagen bezüglich des Durchlaufs von Anforderungen gemacht werden.

3.1.2 Behandelte Probleme

Ziel dieses Kapitels 3 ist es, die Abhängigkeiten zwischen den zahlreichen "Schicksalsgrößen" einer Anforderung (Wartezeiten etc) in den aufeinanderfolgenden Stufen zu untersuchen und außerdem bestehende Lösungen zu verallgemeinern (vgl. 2.2.2.2).
Hierzu werden mehrere Test-Anforderungen betrachtet mit mehr oder weniger bekanntem Schicksal in Stufe 1.
Dies sind zunächst Anforderungen mit speziellen Annahmen über deren Wartezeiten in Stufe 1 (kein Warten, Warten von unbekannter Dauer, bekannte Wartezeit >0). Durch Bestimmung der Zahl der von diesen Test-Anforderungen bei Ankunft in Stufe 2 angetroffenen Anforderungen wird der Einfluß dieser verschiedenen Größen auf das zukünftige Schicksal in Stufe 2 bestimmt. Dabei werden sowohl Anforderungen mit bekannter als auch unbekannter Bedienungszeit in Stufe 1 betrachtet und so der Einfluß der Bedienungszeit in Stufe 1 gezeigt (Abschnitt 3.2).
Die Kenntnis all dieser Abhängigkeiten ist die Vorbedingung für die Bestimmung von Gesamtschicksalen (VF der Gesamtwartezeit inclusive Gesamt-Wartewahrscheinlichkeit und mittlere Gesamtwartezeit der Wartenden). Da in einem vorangestellten Abschnitt 3.2.1 gezeigt wird, daß die zusätzliche Annahme eines konkreten Wertes >0 der Wartezeit in Stufe 1 keinen Einfluß besitzt ("begrenzte Abhängigkeit" der Wartezeiten genannt), ist es möglich, die VF der Gesamtwartezeit durch Faltung spezieller Ausdrücke zu gewinnen (Abschnitt 3.3).

Im letzten Abschnitt 3.4 wird eine weitere Gruppe von Test-Anforderungen betrachtet,nämlich solche,die eine bestimmte Zahl von Anforderungen in Stufe 1 vorgefunden haben.Durch sog. RANDOM WALK-Betrachtungen wird die Zahl der jeweils angetroffenen Anforderungen in der 2.(Single-Server) Stufe bestimmt.Es wird gezeigt,daß für diese Antreffwahrscheinlichkeiten keine solche nur "begrenzte Abhängigkeit" existiert wie bei den Wartezeiten. Es sie hier noch vermerkt,daß diese "begrenzte Abhängigkeit" jedoch nur für $n_1=1$ existiert.

3.2 Test-Anforderungen mit bekannter Wartezeit in Stufe 1

3.2.1 M|M|n-Ausgangsprozeß während der Wartezeit einer Test-Anforderung

Ziel dieser Voruntersuchung ist es,Aussagen zu erhalten über den Ausgangsprozeß eines stationären einstufigen Wartesystems M|M|n mit FIFO Abfertigungsdisziplin während einer bekannten Wartezeit $T_W=t_o(>0)$ einer Test-Anforderung.Es sei

$$p_w(j,t_o) \overset{def}{=} P(X_A=j \mid T_W=t_o)$$

die Wahrscheinlichkeit,daß diese Test-Anforderung bei ihrer Ankunft genau j ($\geq n$) Anforderungen in der ganzen Stufe angetroffen hat.Wendet man das bekannte Theorem von BAYES (vgl.|9|) für bedingte Wahrscheinlichkeiten an,so erhält man

$$P(X_A=j \mid T_W\in[t_o,t_o+dt]) = \frac{P(X_A=j)}{P(T_W\in[t_o,t_o+dt])} \cdot P(T_W\in[t_o,t_o+dt] \mid X_A=j) \qquad (3.9)$$

Die Ausdrücke auf der rechten Seite lauten wie folgt:

$$P(X_A=j) = (1-\varrho)\cdot\varrho^{j-n} \qquad (j\geq n) \qquad (3.10)$$

ist die Wahrscheinlichkeit,daß eine wartende Anforderung das System im Zustand $j\geq n$ angetroffen hat.

$$P(T_W\in[t_o,t_o+dt]) \approx (\mu-\lambda)\cdot e^{-(\mu-\lambda)t_o}\cdot dt \qquad (3.11)$$

entspricht der mit dt multiplizierten Wahrscheinlichkeitsdichte der Wartezeit-VF der Wartenden an der Stelle $t=t_o$,während

$$P(T_W\in[t_o,t_o+dt] \mid X_A=j) \approx \mu\cdot e^{-\mu t_o}\cdot\frac{(\mu t_o)^{j-n}}{(j-n)!}\cdot dt \qquad (3.12)$$

mit Hilfe der Faltung von j-n+1 negativ-exponentiellen Phasen mit Rate $\mu = n\cdot\xi$ dargestellt werden kann.
Durch Einsetzen von (3.10-12) in (3.9) und den Grenzübergang $dt\to 0$ erhält man

$$p_w(j,t_o) = e^{-\lambda t_o}\cdot\frac{(\lambda t_o)^{j-n}}{(j-n)!} \qquad t_o>0, j\geq n \qquad (3.13)$$

Es ist offensichtlich,daß dieser Ausdruck (von ξ unabhängig!) identisch ist mit der Wahrscheinlichkeit,daß j-n=z Anforderungen während t_o ankommen.Dies wird formuliert in

SATZ 3.1: Die Wahrscheinlichkeit,daß eine Anforderung mit konkreter Wartezeit $T_W=t_o$ (>0) genau z ($\geq$0) Anforderungen im Wartespeicher der M|M|n-Stufe angetroffen hat,ist gleich der Wahrscheinlichkeit,daß dieselbe Anforderung bei Verlassen des Wartespeichers (um den Service zu beginnen) genau z ($\geq$0) Anforderungen dort zurücklässt.

Ein ähnlicher Satz wurde von BURKE |60| mit einer anderen Berechnungsmethode für die gesamte Stufe aufgestellt,bezog sich also auf eine Anforderung mit fester Durchlaufszeit T_F.Dies wurde "arrival-departure symmetry" der Durchlaufszeiten genannt,die also auch bezüglich der Wartezeiten gilt,solange Ankünfte und Abgänge aus dem Wartespeicher betrachtet werden.
Vergleicht man beide obigen Sätze mit einem 1972 von COOPER |6| diskutierten Theorem über die Gleichheit der Zustandswahrscheinlichkeiten zu Abgangs- und Ankunftszeitpunkten,so folgt letzteres für diesen Fall M|M|n aus beiden Sätzen,aber nicht umgekehrt.

Wenn eine Anforderung mit konkreter Wartezeit $t_o(>0)$ $j(\geq n)$ Anforderungen in der ganzen Stufe angetroffen hat (dies geschieht bei Stationarität mit der Wahrscheinlichkeit $p_w(j,t_o)$),müssen genau j-n+1 Anforderungen (fertig) bedient werden,bis ihre Bedienung beginnt.Da der Abgang der letzten Anforderung von ihnen mit dem Ende der Wartezeit der Dauer t_o zusammenfällt,verlassen genau j-n Anforderungen das System während der Wartezeit selbst, dies geschieht absolut gesehen mit der Wahrscheinlichkeit

$$e^{-\lambda t_o}\cdot\frac{(\lambda\cdot t_o)^{j-n}}{(j-n)!} ,$$

was für jedes feste j ein Poissonprozeß während t_o mit der Rate λ bedeutet,also auch für unbekanntes j ($\geq n$) bei festem t_o.

Damit ergibt sich

SATZ 3.2: Die Annahme oder Kenntnis eines konkreten Wertes der Wartezeit einer Anforderung im System M|M|n führt zu keiner näheren Aussage über den Ausgangsprozeß während dieser Wartezeit, der deshalb ebenfalls ein Poissonprozeß ist mit der Rate λ.

Zur Abrundung sei vermerkt, daß bei Annahme einer Test-Anforderung, die eine feste Zahl j von Anforderungen angetroffen hätte, der Ausgangsprozeß während der (nun unbekannten) Wartezeit der Anforderung selbstredend auch durch einen Poissonprozeß hätte beschrieben werden können, jedoch mit der Rate $\mu = n \cdot \xi$!
Ähnliche Untersuchungen in dieser Richtung wurden von BURKE |60| vorgenommen. Dieser betrachtete jedoch sog. Teilwartezeiten ("partial delays") für Anforderungen, die mindestens γ-1 Anforderungen im Wartespeicher vorfinden, bis sie den Warteplatz Nr. γ (γ>0) verlassen.
Es kann hier noch bemerkt werden, daß im Falle n_1=1 mit genau derselben Methode wie hier für die Wartezeiten ein alternativer Beweis der Unabhängigkeit des Schicksals einer Anforderung in Stufe 2 von ihrer Durchlaufszeit in Stufe 1 erhalten werden kann.

3.2.2 Allgemeine Berechnungsmethode

Für jede der verschiedenen Test-Anforderungen -die jeweils für sich im System sei und in die betrachtete Richtung gehen soll - wird die Zahl der bei Ankunft in Stufe 2 angetroffenen Anforderungen bestimmt. Wegen der Gedächtnislosigkeit der negativ-exponentiellen VF können alle weiteren Schicksalsgrößen hieraus berechnet werden.
Die Beobachtung des Systems beginne zu Zeit T, wenn die feste Bedienungszeit $T_{H1}=t_1$ einer Test-Anforderung beginnt.

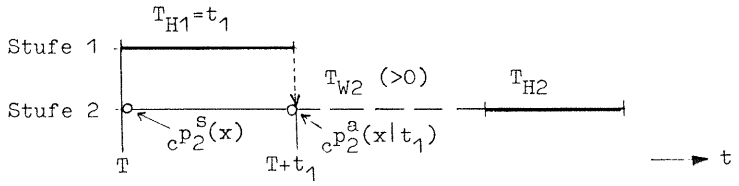


Bild 3.3 Zeitdiagramm zur Berechnung

Aus den Zustandswahrscheinlichkeiten $_cp_2^s(x)$ der Stufe 2 zur Zeit T^+ (Startwahrscheinlichkeit), die unabhängig von der geforderten Bedienungszeit t_1 der Test-Anforderung ist, werden die Zustandswahrscheinlichkeiten $_cp_2^a(x|t_1)$ zur Zeit $(T+t_1)^-$ bestimmt (Antreffwahrscheinlichkeiten). Der vorangestellte Index c gibt an, daß die Wahrscheinlichkeit eine bedingte Wahrscheinlichkeit ist (conditional probability) und sich auf eine spezielle Anforderung bezieht, wobei

- c=0 eine Anforderung mit $T_{W1}=0$ und
- c=1 eine Anforderung mit $T_{W1}>0$ von bekannter oder unbekannter Dauer

bedeutet.
Die so erhaltenen Ergebnisse sind gültig für eine bestimmte Bedienungszeit in Stufe 1. Durch Integration über alle möglichen Bedienungszeiten erhält man die jeweiligen Ergebnisse für Anforderungen mit unbekannter Bedienungszeit T_{H1}.

3.2.3 Startwahrscheinlichkeiten in Stufe 2

Mit den Theoremen von R.R.P. bzw. J.R. JACKSON (vgl. 2.2.2.1 bzw. 2.4.3) ist es offensichtlich, daß die Startwahrscheinlichkeiten $_op_2^s(x)$ in Stufe 2 bei der Ankunft einer nichtwartenden Anforderung in Stufe 1 unabhängig und gemäß den absoluten Zustandswahrscheinlichkeiten für Stufe 2 sind:

$$_op_2^s(x) = p_2(x) \qquad (3.14)$$

wobei

$$p_2(x) = \begin{cases} p_2(0) \cdot \dfrac{A_2^x}{x!} & 0 \le x \le n_2 \\ p_2(0) \cdot \dfrac{A_2^x}{n_2! \cdot n_2^{x-n_2}} & x \ge n_2 - 1 \end{cases} \qquad (3.15)$$

mit

$$p_2(0) \cdot \frac{A_2^{n_2}}{n_2!} \cdot \frac{1}{1 - \frac{A_2}{n_2}} = E_{2,n2}(A_2) = W_2 \qquad (3.16)$$

als

als absoluter Wartewahrscheinlichkeit in Stufe 2 nach der "2. ERLANG'schen Formel", vgl. etwa SYSKI |19|.
Für Anforderungen, die in der 1. Stufe warten müssen, liegt folgende Situation vor:

Zur Zeit $T-t_o$ kommt die betrachtete Test-Anforderung mit Wartezeit $t_o(>0)$ in Stufe 1 an. Nun sind die Zustandswahrscheinlichkeiten von Stufe 2 zur Zeit $T-t_o$ wiederum identisch mit den absoluten Werten nach (3.15). Während der folgenden Zeit t_o ist der Eingangsprozeß in Stufe 2 ein Poissonprozeß (gezeigt in 3.2.1, Verzweigung unerheblich), deshalb sind die Zustandswahrscheinlichkeiten der 2. Stufe zur Zeit T^- auch mit den absoluten identisch. Damit gilt zur Zeit T^+ für die Startwahrscheinlichkeiten

$$ {}_1p_2^s(x) = \begin{cases} (1-\eta)\cdot p_2(x) & \text{für } x=0 \\ \eta\cdot p_2(x-1) + (1-\eta)\cdot p_2(x) & \text{für } x>0 \end{cases} \qquad (3.17) $$

Es ist klar, daß (3.17) auch gilt, wenn der konkrete Wert der Wartezeit unbekannt ist oder wenn eine bestimmte Bedienungszeit $T_{H1}=t_1$ einer Anforderung zugewiesen wird.
Vergleicht man (3.14) mit (3.17), so erkennt man z.B. für $\eta=1$ (ohne Verzweigung) einen Unterschied von genau einer Anforderung.

3.2.4 Antreffwahrscheinlichkeiten in Stufe 2

Es sei $p(i,t_1)$ die Wahrscheinlichkeit, daß i Anforderungen während der Bedienungszeit $T_{H1}=t_1$ einer Test-Anforderung Stufe 2 verlassen. So lange wie Stufe 2 voll beschäftigt ist ($X_2 \geq n_2$) so lange ist die Rate der ganzen Stufe $\mu_2 = n_2\cdot\varepsilon_2$ (vgl. Bild 3.3). Wenn während T_{H1} X_2 kleiner n_2 wird, haben BE der Stufe 2 Freizeiten, was eine momentan geringere Ausgangsrate bedeutet. Aber in diesen Fällen muß die Test-Anforderung in Stufe 2 nicht warten. Deshalb lauten die Antreffwahrscheinlichkeiten

$$ {}_cp_2^a(x|t_1) = \sum_{i=0}^{\infty} p(i,t_1)\cdot {}_cp_2^s(x+i) \qquad x\geq n_2,\ c=0,1, \qquad (3.18) $$

wobei

$$ p(i,t_1) = e^{-\mu_2 t_1}\cdot\frac{(\mu_2 t_1)^i}{i!} \qquad i\geq 0 \qquad (3.19) $$

die Wahrscheinlichkeit für i Poissonereignisse mit Parameter μ_2 während der Zeitdauer t_1 darstellt.
Durch Integration erhält man die Antreffwahrscheinlichkeiten für entsprechende Test-Anforderungen mit unbekanntem T_{H1}:

$$ {}_cp_2^a(x) = \int_{t_1=0}^{\infty} {}_cp_2^a(x|t_1)\cdot f_{H1}(t_1)\,dt_1 \qquad c=0,1. \qquad (3.20) $$

Für Test-Anforderungen, die in Stufe 1 nicht gewartet haben, erhält man aus (3.18) mit (3.19) und ${}_op_2^s(x+i)$ gemäß (3.14) aus (3.15) mit (3.16) für $x\geq n_2$

$$ P(X_{A2}=x|T_{W1}=0, T_{H1}=t_1) \overset{def}{=} {}_op_2^a(x|t_1) = p_2(x)\cdot e^{-(\mu_2-\lambda_2)t_1} \qquad (3.21) $$

Falls die Annahme einer bestimmten Bedienungszeit fallengelassen wird, erhält man aus (3.20) mit (3.2)

$$ P(X_{A2}=x|T_{W1}=0) \overset{def}{=} {}_op_2^a(x) = \frac{\varepsilon_1}{\varepsilon_1+\mu_2-\lambda_2}p_2(x) \qquad x\geq n_2 \qquad (3.22) $$

Auf gleiche Weise erhält man die Antreffwahrscheinlichkeiten für Anforderungen mit $T_{W1}>0$ von bekannter oder unbekannter Dauer (c=1)

$$ \left.\begin{aligned} {}_1p_2^a(x|t_1) &= \left[\eta\cdot p_2(x-1)+(1-\eta)p_2(x)\right]\cdot e^{-(\mu_2-\lambda_2)t_1} \\ &= p_2(x-1)\left[\eta+(1-\eta)\cdot\frac{A_2}{n_2}\right]\cdot e^{-(\mu_2-\lambda_2)t_1} \end{aligned}\right\} x\geq n_2 \qquad (3.23) $$

$$ {}_1p_2^a(x) = \frac{\varepsilon_1}{\varepsilon_1+\mu_2-\lambda_2}p_2(x-1)\cdot\left[\eta+(1-\eta)\cdot\vartheta_2\right] \qquad (3.24) $$

Für Anforderungen, bei denen nur die geforderte oder verflossene Bedienungszeit T_{H1} bekannt ist, folgt aus (3.21) und (3.23) durch gewichtete Summation

$$ P(X_{A2}=x|T_{H1}=t_1) \overset{def}{=} p_2^a(x|t_1) = (1-\eta\cdot\vartheta_1+\frac{\eta\vartheta_1}{\vartheta_2})p_2(x)e^{-(\mu_2-\lambda_2)t_1} \qquad x\geq n_2 \qquad (3.25) $$

3.2.5 Bedingte Wartewahrscheinlichkeiten

Durch Aufsummation aller Antreffwahrscheinlichkeiten für $x\geq n_2$ erhält man nun die Wartewahrscheinlichkeiten für die verschiedenen Test-Anforderungen

$$ {}_cP(T_{W2}>0|T_{H1}=t_1) = \sum_{x=n_2}^{\infty} {}_cp_2^a(x|t_1) \qquad c=0,1, \qquad (3.26) $$

bzw. falls die Bedienungszeit unbekannt ist

$$ {}_cW_2 = \sum_{x=n_2}^{\infty} {}_cp_2^a(x) \qquad c=0,1. \qquad (3.27) $$

(3.26) liefert mit (3.21) und (3.23) nach einigen Zwischenrechnungen

$$ P(T_{W2}>0 \mid T_{W1}=0, T_{H1}=t_1) \overset{def}{=} {}_oP(T_{W2}>0 \mid T_{H1}=t_1) = E_{2,n2}(A_2)\cdot e^{-(\mu_2-\lambda_2)t_1} \qquad (3.28) $$

und

$$ {}_1P(T_{W2}>0 \mid T_{H1}=t_1) = \left(\frac{\eta}{\vartheta_2}+1-\eta\right)\cdot E_{2,n2}(A_2)\; e^{-(\mu_2-\lambda_2)t_1} \qquad (3.29) $$

In beiden Fällen sinkt die Wartewahrschkt. in Stufe 2 exponentiell mit der Bedienungszeit $T_{H1}=t_1$; je größer T_{H1}, desto mehr Anforderungen können währenddessen in Stufe 2 abgefertigt werden. Für Anforderungen mit unbekanntem T_{H1} erhält man

$$ P(T_{W2}>0 \mid T_{W1}=0) \overset{def}{=} {}_oW_2 = {}_oF\cdot E_{2,n2}(A_2) $$

mit

$$ {}_oF = \frac{\xi_1}{\xi_1+\mu_2-\lambda_2} = \frac{\vartheta_2}{\eta\vartheta_1+\vartheta_2-\eta\vartheta_1\vartheta_2} \qquad (<1) \qquad (3.30) $$

und

$$ P(T_{W2}>0 \mid T_{W1}>0) \overset{def}{=} {}_1W_2 = {}_1F\cdot E_{2,n2}(A_2) $$

mit

$$ {}_1F = \frac{\xi_1\cdot\left(\frac{\eta}{\vartheta_2}+1-\eta\right)}{\xi_1+\mu_2-\lambda_2} = \frac{\eta+(1-\eta)\cdot\vartheta_2}{\eta\cdot\vartheta_1+\vartheta_2-\eta\vartheta_1\vartheta_2} \qquad (>1) \qquad (3.31) $$

Vergleicht man jeweils die bedingten Wartewahrscheinlichkeiten (3.28) und (3.29) bzw. (3.30) und (3.31), so gilt

$$ P(T_{W2}>0 \mid T_{W1}=0, T_{H1}=t_1) = F(\eta,\vartheta_2)\cdot P(T_{W2}>0 \mid T_{W1}>0, T_{H1}=t_1) \qquad (3.32) $$

und

$$ {}_oW_2 = F(\eta,\vartheta_2)\cdot {}_1W_2 \qquad (3.33) $$

wobei der Relationsfaktor

$$ F(\eta,\vartheta_2) = \frac{1}{\frac{\eta}{\vartheta_2}+1-\eta} = \frac{P(T_{W2}>0 \mid T_{W1}=0)}{P(T_{W2}>0 \mid T_{W1}>0)} \qquad (3.34) $$

beträgt und in Bild 3.4 dargestellt ist.
Für $\eta=1$ (keine Verzweigung) ist der Unterschied zwischen beiden bedingten Wartewahrscheinlichkeiten am größten, während für $\eta\to 0$ der Unterschied verschwindet.

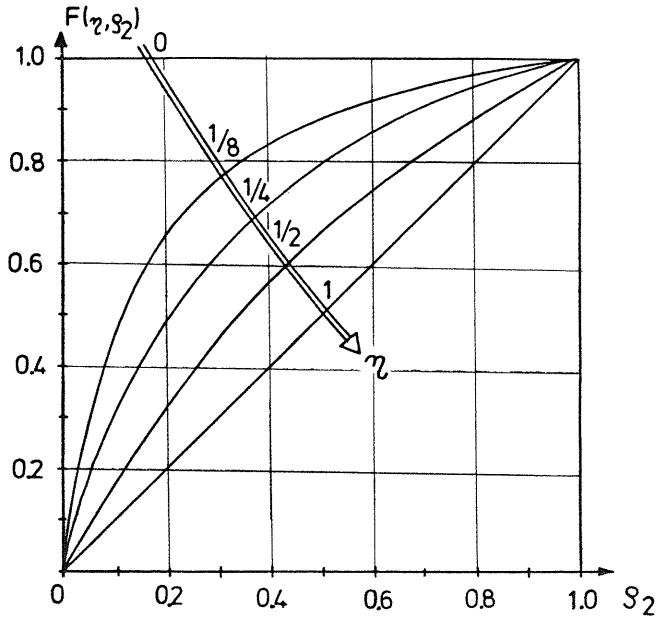


Bild 3.4 Relationsfaktoren $F(\eta,\vartheta_2)$

Speziell für $\eta=1$, $n_2=1$ vereinfachen sich (3.30) und (3.31) zu denselben Ergebnissen, die von BURKE [59] abgeleitet wurden.
Die bedingten Wartewahrscheinlichkeiten für unbekanntes T_{H1} sind für $\eta=1$ in den Bildern 3.5 und 3.6 aufgetragen über der Ausnutzung $\vartheta_2 = A_2/n_2$ der Stufe 2 mit ϑ_2/ϑ_1 als Parameter.

Alle Kurven für ${}_oW_2$ verlaufen unterhalb $W_2 = E_{2,n2}(A_2)$, alle Kurven für ${}_1W_2$ oberhalb.
Mit Hilfe des Grenzwertes

$$ \lim_{A_2\to 0} \frac{E_{2,n2}(A_2)}{\frac{A_2}{n_2}} = \begin{cases} 1 & \text{für } n_2=1 \\ 0 & \text{für } n_2>1 \end{cases} $$

kann man zeigen, daß sich bei $n_2=1$ bei verschwindendem Angebot ein endlicher Wert

$$ \lim_{A_2\to 0} P(T_{W2}>0 \mid T_{W1}>0) = \frac{\eta}{\frac{h_1}{h_2}+1} \qquad (3.35) $$

ergibt, welcher sich plausibel erklären ließe.

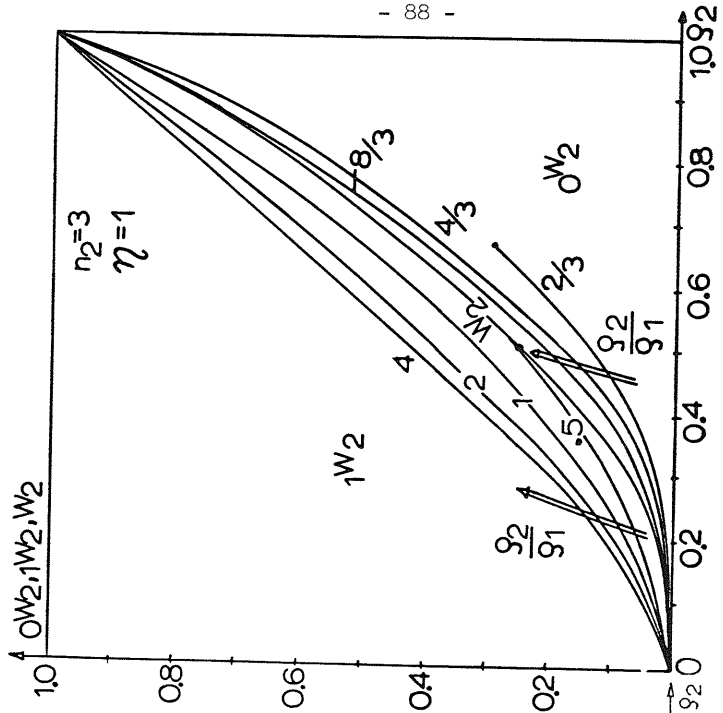


Bild 3.6 Bedingte Wartewahrscheinlichkeiten in Stufe 2 ($n_2=3$)

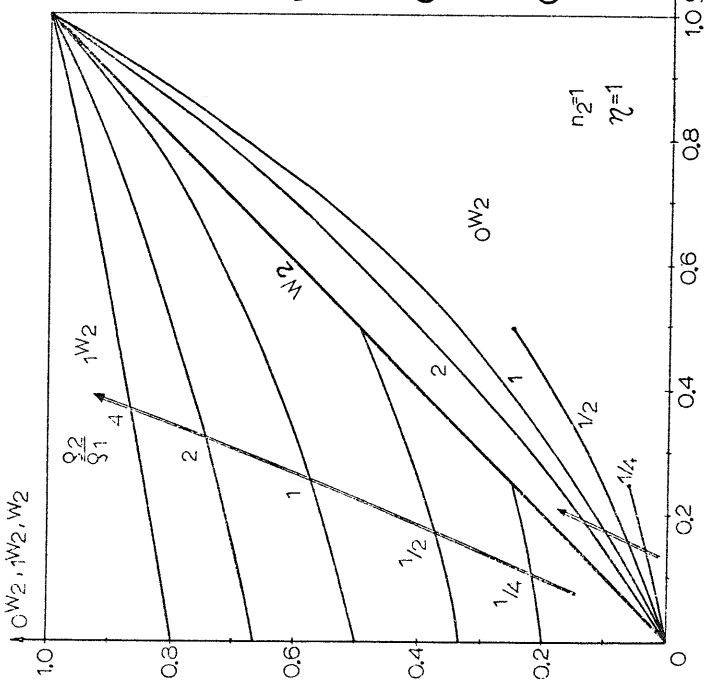


Bild 3.5 Bedingte Wartewahrscheinlichkeiten in Stufe 2 ($n_2=1$)

3.2.6 Bedingte Wartezeitverteilungsfunktionen

Wird angenommen, daß eine Test-Anforderung in der 2.Stufe warten muß, so ergeben sich durch Normierung der Antreffwahrscheinlichkeiten für $x \geqslant n_2$ auf die entsprechenden bedingten Wartewahrscheinlichkeiten für jede der betrachteten Test-Anforderungen die gleichen bedingten Antreffwahrscheinlichkeiten wie für alle wartenden Anforderungen in Stufe 2 zusammen:

$$P(X_{A2}=x \mid T_{W2}>0, \ldots) = (1-\frac{A_2}{n_2})\cdot(\frac{A_2}{n_2})^{x-n_2} \qquad x \geqslant n_2. \qquad (3.36)$$

Deshalb gilt folgender Satz

SATZ 3.3: Sobald eine Anforderung in Stufe 2 warten muß, ist die Zahl der angetroffenen Anforderungen und damit -bei FIFO Abfertigungsdisziplinen in beiden Stufen - ihr zukünftiges Schicksal (Wartezeit, Durchlaufszeit) unabhängig vom vergangenen Schicksal (Wartezeit, Bedienungszeit, Durchlaufszeit) in Stufe 1.

Mit den Ergebnissen aus 3.4 kann man zeigen, daß dies nur gilt, wenn keine feste Zahl angetroffener Anforderungen X_{A1} (>0) in Stufe 1 angenommen wird.

Für FIFO Abfertigungsdisziplin in Stufe 2 hat jede solche Test-Anforderung die gleiche bedingte (komplementäre) Wartezeit-VF wie alle wartenden Anforderungen in Stufe 2 zusammen:

$$P(T_{W2}>t \mid T_{W2}>0) \overset{def}{=} W_{w2}(>t) = e^{-(\mu_2-\lambda_2)t} \qquad (3.37)$$

Ohne die Bedingung des Wartens in Stufe 2 ergibt sich aus (3.37) mit (3.28)-(3.31)

$${}_cP(T_{W2}>t \mid T_{H1}=t_1) = (\frac{\eta}{\varrho_2}+1-\eta)^c \cdot E_{2,n2}(A_2)\cdot e^{-(\mu_2-\lambda_2)\cdot(t_1+t)} \qquad (3.38)$$

$$c=0,1.$$

$${}_cW_2(>t) = {}_cF\cdot E_{2,n2}(A_2)\cdot e^{-(\mu_2-\lambda_2)t} \qquad (3.39)$$

Aus (3.38) erhält man durch gewichtete Summation die bedingte Wartezeit-VF für Anforderungen, bei denen nur ihre Bedienungszeit in Stufe 1 bekannt ist:

$$P(T_{W2}>t \mid T_{H1}=t_1) = P(T_{W2}>t, T_{W1}=0 \mid T_{H1}=t_1) + P(T_{W2}>t, T_{W1}>0 \mid T_{H1}=t_1)$$

$$= (1-\eta\varrho_1+\frac{\eta\varrho_1}{\varrho_2})\cdot E_{2,n2}(A_2)\, e^{-(\mu_2-\lambda_2)(t_1+t)} \qquad (3.40)$$

Für Anforderungen mit $T_{H1}=t_1 \to 0$ ergibt sich hieraus nicht die absolute Wartezeit-VF in Stufe 2, weil diese Anforderungen möglicherweise in Stufe 1 warten müssen.
Für alle angegebenen bedingten Wartezeit-VF lassen sich auch in einfacher Weise die zugehörigen bedingten Erwartungswerte für T_{W2} bestimmen; Ziel der Untersuchungen ist jedoch das Gesamtschicksal von Anforderungen, das nun bestimmt werden kann.

3.3 Bestimmung von Gesamtschicksalen

Die Unabhängigkeit der Durchlaufszeiten im betrachteten System bewirkt, daß die VF der Gesamtdurchlaufszeit einfach als Faltung der beiden Durchlaufszeit-VF der Einzelstufen berechnet werden kann.
Bei der Bestimmung der VF der Gesamtwartezeit müssen die Abhängigkeiten der Wartezeiten T_{W1} und T_{W2} berücksichtigt werden. Dies ist nun möglich, nachdem in 3.2 nicht nur gezeigt wurde, daß sie existieren, sondern auch vollständig bestimmt wurden.

3.3.1 Gesamtwartewahrscheinlichkeit

Die Gesamtwartewahrscheinlichkeit W ist die Wahrscheinlichkeit, daß eine Anforderung irgendwo im System warten muß:

$$\begin{aligned} W = P(T_W>0) &= P(T_{W1}>0)+P(T_{W1}=0)\cdot P(T_{W2}>0 \mid T_{W1}=0) \\ &= P(T_{W1}>0)+P(T_{W2}>0)-P(T_{W1}>0, T_{W2}>0) \end{aligned} \qquad (3.41)$$

Benutzt man (3.30) oder (3.31), so kann man folgende einfache Form für W finden:

$$W = W_1 + W_2 - \frac{(1-\eta)S_2 + \eta}{\eta S_1 + S_2 - \eta S_1 S_2} \cdot W_1 \cdot W_2$$

wobei

$$W_1 = S_1 = \frac{\lambda}{\varepsilon_1} \;,\; W_2 = E_{2,n2}(A_2) \;,\; S_2 = \frac{A_2}{n_2} = \frac{\lambda_2}{n_2 \varepsilon_2} \qquad (3.42)$$

In Bild 3.7 ist die Gesamtwartewahrscheinlichkeit gezeigt für den einfachsten Fall $\eta=1$ und $n_2=1$.

Zum Vergleich sind die bei Annahme der Unabhängigkeit erhaltenen Werte

$$W_{appr} = W_1 + W_2 - W_1 W_2 \qquad (3.43)$$

gestrichelt eingezeichnet.

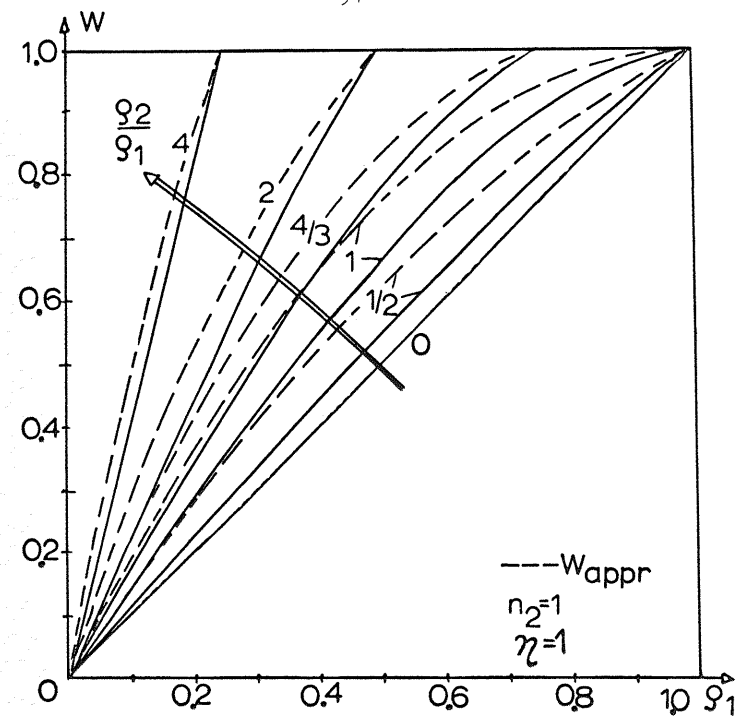


Bild 3.7 Gesamtwartewahrscheinlichkeit für $\eta=1, n_2=1$ (hier $\frac{S_2}{S_1}=\frac{h_2}{h_1}$)

Die Gesamtwartewahrscheinlichkeit ist immer kleiner als der durch Annahme der Unabhängigkeit erhaltene Wert, weil solche Anforderungen in Stufe 2 bevorzugt warten, die bereits in Stufe 1 warten mußten.
Die einfache exakte Lösung nach (3.42) macht nunmehr die Approximation nach (3.43) entbehrlich.
In obigem dargestellten Falle mit $\eta=1$, $n_2=1$ erhält man eine von der Reihenfolge der beiden Stufen unabhängige Gesamtwartewahrscheinlichkeit ($\lambda_1=\lambda_2=\lambda$)

$$W = \frac{\lambda}{\varepsilon_1} + \frac{\lambda}{\varepsilon_2} - \frac{\lambda}{\varepsilon_1+\varepsilon_2-\lambda} \qquad (3.44)$$

Bestimmt man die Gesamtwartewahrscheinlichkeit für Anforderungen mit bestimmter Bedienungszeit in Stufe 1 mit einer (3.41) analogen Beziehung, so ergibt sich mit (3.28) ein Ausdruck für die Abhängigkeit dieser Größe von der Bedienungszeit $T_{H1}=t_1$:

$$P(T_W>0 \mid T_{H1}=t_1) = W_1+(1-W_1)\cdot W_2 \cdot e^{-(\mu_2-\lambda_2)t_1} \qquad (3.45)$$

3.3.2 Mittlere Gesamtwartezeit der Wartenden

Die mittlere Gesamtwartezeit aller Anforderungen ist die Summe der mittleren Wartezeiten aller Anforderungen in den Einzelstufen und deshalb unabhängig von der Reihenfolge der Stufen (dies gilt,wie hier,auch bei abhängigen Wartezeiten):

$$E(T_W) = E(T_{W1}) + E(T_{W2}) \tag{3.46}$$

mit

$$E(T_{Wi}) = W_i \cdot t_{Wi} \tag{3.47}$$

wobei

$$t_{Wi} = \frac{1}{n_i - A_i} \cdot h_i = \frac{1}{\mu_i - \lambda_i} \qquad i=1,2, \tag{3.48}$$

die mittlere Wartezeit der Wartenden in Stufe i ist.
Für die mittlere Gesamtwartezeit der Wartenden gilt allgemein

$$t_W = \frac{E(T_W)}{W} \quad , \tag{3.49}$$

sinngemäß auch für spezielle Test-Anforderungen.
Sind z.B.2 gleiche Single-Server Stufen in Serie ohne Verzweigung ,so ergibt sich ($A_1=A_2=A, h_1=h_2=h$)

$$t_W = \frac{2-A}{(1-A)\cdot(3-2A)} \cdot 2h \tag{3.50}$$

3.3.3 Verteilungsfunktion der Gesamtwartezeit

In einem 2-stufigen System gilt (unabhängige Verzweigung zugelassen) allgemein

$$P(T_W>t)=P(T_{W2}=0)\cdot P(T_{W1}>t|T_{W2}=0)+P(T_{W1}=0)\cdot P(T_{W2}>t|T_{W1}=0)+$$
$$+P(T_{W1}>0,T_{W2}>0)\cdot P(T_{W1}+T_{W2}>t|T_{W1}>0,T_{W2}>0) \tag{3.51}$$

In dieser gewichteten Summation (komplementärer) VF sind alle Wichtungswahrscheinlichkeiten nun bekannt.
Weil der konkrete Wert von $T_{W1}(>0)$ keinen Einfluß besitzt (vgl. (vgl. 3.2.1),kann gezeigt werden,daß

$$P(T_{W1}>t|T_{W2}=0) = P(T_{W1}>0|T_{W2}=0)\cdot P(T_{W1}>t|T_{W1}>0) \tag{3.52}$$

$P(T_{W2}>t|T_{W1}=0)$ wurde in (3.39) angegeben.Der letzte Term in (3.51) bezieht sich auf Anforderungen,die in beiden Stufen warten.Da gezeigt wurde,daß in dem betrachteten System die Zufallsvariablen T_{W1} und T_{W2} nur dann als voneinander unabhängig angesehen werden können,wenn beide >0 angenommen werden,ist

unter dieser Nebenbedingung Faltung erlaubt:

$$P(T_{W1}+T_{W2}>t|T_{W1}>0,T_{W2}>0) = P(T_{W1}>t|T_{W1}>0) \ast P(T_{W2}>t|T_{W2}>0)$$
$$= e^{-(\varepsilon_1-\lambda_1)t} \ast e^{-(\mu_2-\lambda_2)t} \tag{3.53}$$

Hiermit läßt sich die VF der Gesamtwartezeit für beliebige in Richtung der betrachteten Stufe 2 gehenden Anforderungen bestimmen,als auch analog für Anforderungen mit bestimmter Bedienungszeit in Stufe 1.Diese Herleitung ist formal gleich,es erscheint in jedem Term $T_{H1}=t_1$ als Zusatzbedingung.Deshalb sei hier nur deren Ergebnis für die VF der Gesamtwartezeit dargestellt:

$$P(T_W>t|T_{H1}=t_1) = \left[1-\left(\frac{\eta}{\vartheta_2}+1-\eta\right)\cdot\frac{\varepsilon_1-\lambda_1}{\varepsilon_1-\lambda_1-\mu_2+\lambda_2}\cdot W_2 e^{-(\mu_2-\lambda_2)t_1}\right]\cdot\vartheta_1 e^{-(\varepsilon_1-\lambda_1)t}$$
$$+\left[(1-\vartheta_1)+\vartheta_1\left(\frac{\eta}{\vartheta_2}+1-\eta\right)\cdot\frac{\varepsilon_1-\lambda_1}{\varepsilon_1-\lambda_1-\mu_2+\lambda_2}\right]\cdot W_2\cdot e^{-(\mu_2-\lambda_2)\cdot(t_1+t)}$$
$$(\varepsilon_1-\lambda_1 \neq \mu_2-\lambda_2) \tag{3.54}$$

Durch eine einfache Integration von $P(T_W>t|T_{H1}=t_1)$ über t_1 bzw. durch Anwendung von (3.51)-(3.53) erhält man die Gesamtwartezeit-VF für alle Anforderungen,die z.B. für Systeme ohne Verzweigung ($\eta=1,\lambda_1=\lambda_2=\lambda$) lautet:

$$P(T_W>t) \overset{def}{=} W(>t) = \begin{cases} \left\{\frac{\lambda}{\varepsilon_1} - \frac{\varepsilon_1-\lambda}{\varepsilon_1+\mu_2-\lambda}\cdot\frac{\mu_2}{\varepsilon_1-\mu_2}\cdot W_2\right\}\cdot e^{-(\varepsilon_1-\lambda)t} + \\ + \frac{\varepsilon_1-\lambda}{\varepsilon_1+\mu_2-\lambda}\cdot\frac{\varepsilon_1}{\varepsilon_1-\mu_2}\cdot W_2\cdot e^{-(\mu_2-\lambda)t} \\ \qquad\qquad \text{für } \varepsilon_1 \neq \mu_2 \qquad (3.55) \\ \left\{\frac{\lambda}{\mu} + \frac{\mu-\lambda}{2\mu-\lambda}\cdot W_2(1+\mu t)\right\}\cdot e^{-(\mu-\lambda)t} \\ \qquad\qquad \text{für } \varepsilon_1=\mu_2=\mu \qquad (3.56) \end{cases}$$

Falls zusätzlich $n_2=1$,so kann eine in ε_1 und ε_2 symmetrische Form erreicht werden:

$$W(>t) = \begin{cases} \frac{\lambda}{\varepsilon_1}\cdot\frac{\varepsilon_2}{\varepsilon_2-\varepsilon_1}\cdot\frac{\varepsilon_2-\lambda}{\varepsilon_1+\varepsilon_2-\lambda}\cdot e^{-(\varepsilon_1-\lambda)t} \\ \frac{\lambda}{\varepsilon_2}\cdot\frac{\varepsilon_1}{\varepsilon_1-\varepsilon_2}\cdot\frac{\varepsilon_1-\lambda}{\varepsilon_1+\varepsilon_2-\lambda}\cdot e^{-(\varepsilon_2-\lambda)t} \\ \qquad\qquad \text{für } \varepsilon_1 \neq \varepsilon_2 \end{cases} \tag{3.57}$$

SATZ 3.4: Bei 2 seriell angeordneten Single-Server Stufen ohne Verzweigung ist die VF der Gesamtwartezeit unabhängig von ihrer Reihenfolge, beide Stufen sind also bezüglich der Gesamtwartezeit vertauschbar.

Diese Ergebnisse für die VF der Gesamtwartezeit gelten nur für $n_1=1$, weil im Falle $n_1>1$ wegen der Überholmöglichkeit in Stufe 1 das Schicksal einer Anforderung in Stufe 2 vom konkreten Wert der Wartezeit T_{W1} abhängt.

3.3.4 Korrelations- und Fehlerbetrachtungen

In diesem Abschnitt werden aus Umfangsgründen nur Anforderungen betrachtet mit unbekannter Bedienungszeit in Stufe 1; für andere Anforderungen gilt nachfolgendes in analoger Weise.
Durch Annahme der Unabhängigkeit der Wartezeiten erhält man eine approximative VF der Gesamtwartezeit durch die Faltung

$$W(>t)_{appr} = P(T_{W1}>t) \ast P(T_{W2}>t)$$

mit

$$P(T_{Wi}>t) = W_i \cdot e^{-(\mu_i-\lambda_i)t} \qquad i=1,2. \tag{3.58}$$

Als einfaches Beispiel gilt für 2 gleiche Single-Server Stufen in Serie ohne Verzweigung ($\lambda_1=\lambda_2=\lambda, h_1=h_2=h, A=\lambda h$)

$$W(>t)_{appr} = A\left\{2-A+A(1-A)\cdot\frac{t}{h}\right\}e^{-(1-A)\frac{t}{h}} \tag{3.59}$$

Die (komplementäre) VF nach (3.56) lautet

$$W(>t) = \frac{A}{2-A}\left\{3-2A+(1-A)\cdot\frac{t}{h}\right\}e^{-(1-A)\cdot\frac{t}{h}} \tag{3.60}$$

In Bild 3.8 sind diese beiden VF, die den gleichen Erwartungswert $E(T_W)$ besitzen, dargestellt.
Man kann nun erkennen, daß die exakten Kurven eine größere Varianz ergeben als die übliche Näherung, was auf die Abhängigkeit der Wartezeiten zurückzuführen ist.
Es gilt

$$Var(T_W) = Var(T_{W1}) + Var(T_{W2}) + 2Cov(T_{W1},T_{W2})$$

mit

$$Cov(T_{W1},T_{W2}) = E(T_{W1}\cdot T_{W2}) - E(T_{W1})\cdot E(T_{W2}) \tag{3.61}$$

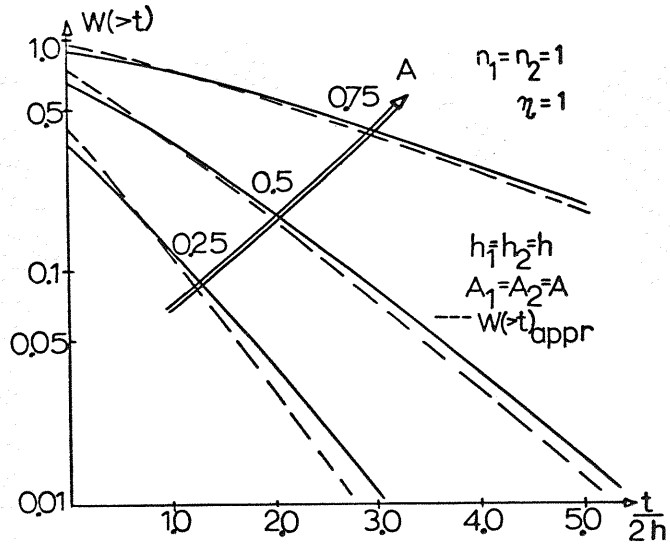


Bild 3.8 Verteilungsfunktion der Gesamtwartezeit

Wegen der "begrenzten Abhängigkeit" von T_{W1} und T_{W2} gilt hierbei

$$E(T_{W1}\cdot T_{W2}) = P(T_{W1}>0, T_{W2}>0)\cdot E(T_{W1}\cdot T_{W2} \mid T_{W1}>0, T_{W2}>0) \tag{3.62}$$

außerdem gilt

$$Var(T_{Wi}) = t_{Wi}^2\cdot W_i\cdot(2-W_i) \qquad i=1,2. \tag{3.63}$$

mit t_{Wi} nach (3.48).
Der Korrelationskoeffizient

$$r(T_{W1},T_{W2}) \overset{def}{=} \frac{Cov(T_{W1},T_{W2})}{\sqrt{Var(T_{W1})}\cdot\sqrt{Var(T_{W2})}}, \tag{3.64}$$

der im Falle $Var(T_{W1})=Var(T_{W2})$ identisch ist mit der relativen Zunahme der Varianz durch Berücksichtigung der (positiven) Kovarianz, kann auf folgenden einfachen Ausdruck zurückgeführt werden:

$$r(T_{W1},T_{W2}) = \frac{{}_1F-1}{\sqrt{\left(\frac{2}{W_1}-1\right)\cdot\left(\frac{2}{W_2}-1\right)}} \tag{3.65}$$

Im Falle $\eta=1$ und $n_2=1$ (vgl. ${}_1F$ nach (3.31)) ergibt sich ein in ε_1 und ε_2 symmetrischer Ausdruck

$$r(T_{W1},T_{W2}) = \frac{1}{\sqrt{(2\varepsilon_1-\lambda)\cdot(2\varepsilon_2-\lambda)}} \cdot \left[\frac{\varepsilon_1\cdot\varepsilon_2}{\varepsilon_1+\varepsilon_2-\lambda} - \lambda \right] \quad , \qquad (3.67)$$

der in Bild 3.9 dargestellt ist als Funktion der Ausnutzungen.

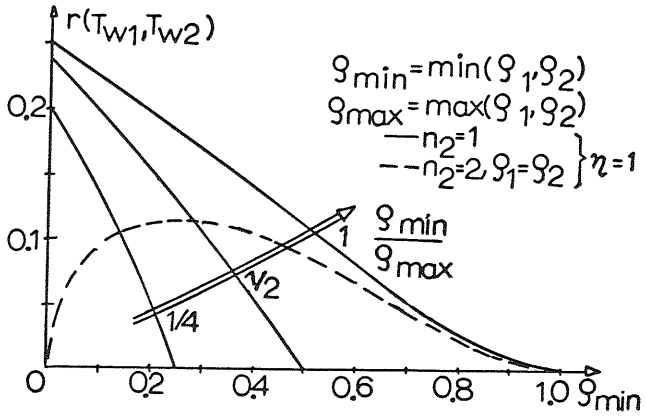


Bild 3.9 Korrelationskoeffizient als Funktion der Ausnutzungen

Zum Vergleich ist noch ein einfaches Beispiel eingezeichnet, das den prinzipiellen Verlauf für $n_2>1$ zeigt. Die Abhängigkeit von der Verzweigungswahrscheinlichkeit η bzw. $1-\eta$ ist in Bild 3.10 dargestellt.

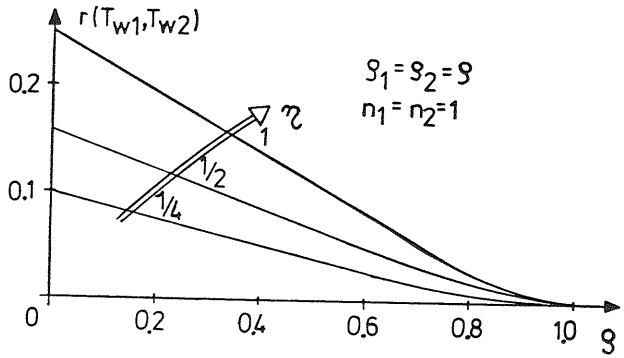


Bild 3.10 Korrelationskoeffizient als Funktion von η

Allgemein kann also gesagt werden, daß die Korrelation sinkt mit zunehmender

- Unsymmetrie in den Wartewahrscheinlichkeiten
- Zahl n_2 der BE in Stufe 2
- Verzweigungswahrscheinlichkeit $1-\eta$

Bestimmt man den relativen Fehler $(W_{appr}-W)/W$, so erhält man prinzipiell dieselben Abhängigkeiten von der Verkehrsintensität wie für die Korrelationskoeffizienten gezeigt wurde. Deshalb seien hier nur einige Maximalwerte dieses relativen Fehlers bei Systemen ohne Verzweigung als Funktion der Zahl n_2 von BE angegeben

n_2	1	2	3	6	10
Fehler	33%	13%	9%	6%	4%

Für $n_2>1$ können diese Fehler infolge der (falschen) Annahme der Unabhängigkeit auftreten, wenn die Ausnutzungen in beiden Stufen so sind, daß die Wartewahrscheinlichkeiten W_1 und W_2 im Bereich 0.2 bis 0.3 liegen.

3.4 Test-Anforderungen mit bestimmter Startposition in Stufe 1

In diesem Abschnitt werden die bedingten Antreffwahrscheinlichkeiten

$$P(X_{A2}=x \mid X_{A1}=x_1) \overset{def}{=} p_2(x|x_1)$$

bestimmt, also die Abhängigkeit der Zahl der Anforderungen (Warteschlangenlängen) die von derselben Anforderung in den beiden Stufen dort angetroffen werden. Um explizite Ergebnisse zu erhalten, wurde der (wichtige) Fall zweier Single-Server Systeme ohne Verzweigung betrachtet, jedoch ist bei Berücksichtigung derselben bzw. auch bei $n_2>1$ ein ähnliches RANDOM WALK-Diagramm zu erhalten, das z.B. rekursiv gelöst werden müsste. Jedoch gelten die im betrachteten Falle erhaltenen prinzipiellen Ergebnisse in der Tendenz auch dort.

3.4.1 Allgemeine Berechnungsmethode

Der Durchlauf einer Anforderung durch das System kann beschrieben werden durch eine Folge von "Durchlaufszuständen", die eine Anforderung annimmt (Weg in einem sog. RANDOM WALK-Diagramm). Da das Schicksal einer Anforderung wegen FIFO-Abfertigungsdisziplin in beiden Stufen nicht von nachfolgenden Anforderungen beeinflußt wird, ist das RANDOM WALK-Diagramm ein gerichteter Graph ohne Schleifen (vgl. Bild 3.11).

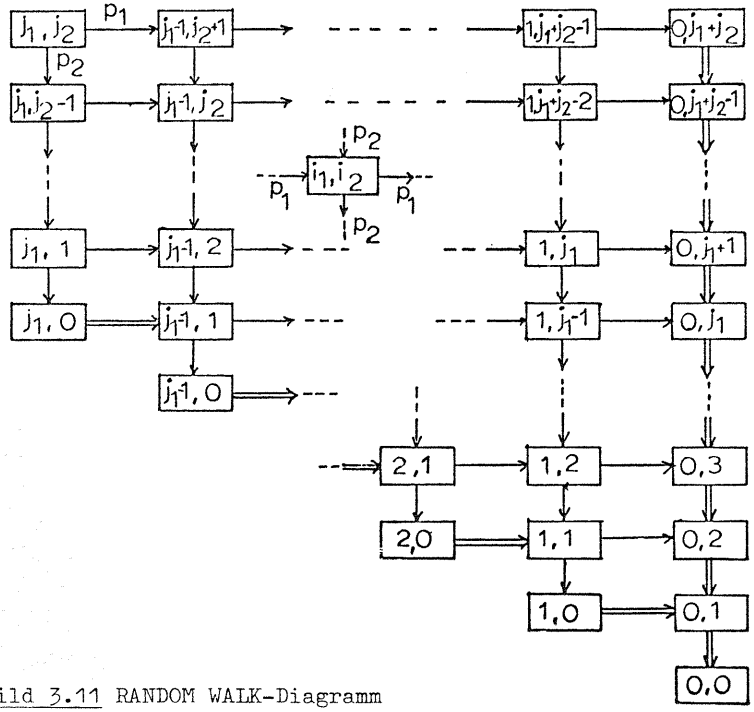


Bild 3.11 RANDOM WALK-Diagramm

Ein allgemeiner Durchlaufzustand $\boxed{i_1,i_2}$ für eine betrachtete Anforderung ist so definiert, daß i_1 Anforderungen in Stufe 1 und i_2 in Stufe 2 sich befinden, inclusive der betrachteten Anforderung, wobei aber nachfolgende Anforderungen irrelevant sind. Wird dieser Durchlaufszustand von der betrachteten Anforderung eingenommen, so ist das nächste Ereignis entweder das Enden einer Bedienungszeit in Stufe 1 oder in Stufe 2 mit den Wahrscheinlichkeiten

$$p_1 = \frac{\varepsilon_1}{\varepsilon_1+\varepsilon_2} \quad ; \quad p_2 = \frac{\varepsilon_2}{\varepsilon_1+\varepsilon_2} \tag{3.68}$$

Ist eine Stufe leer, so zeigen doppelt eingezeichnete Pfeile an, daß die Übergangswahrscheinlichkeit =1 ist (reflektierende Grenze).

Jede Anforderung beginnt ihren Durchlauf beim Start-Durchlaufszustand $\boxed{j_1,j_2}$ und wandert auf einem bestimmten Weg mit zugehöriger Wahrscheinlichkeit zu dem absorbierenden Durchlaufszustand $\boxed{0,0}$, wo die Anforderung das System verläßt.

Es ist nicht schwierig, die Aufenthaltszeit einer Anforderung in einem Durchlaufszustand anzugeben, deshalb könnte man zeigen, wie T_{W1}, T_{H1}, T_{W2} und T_{H2} sich in diesem Diagramm widerspiegeln.

Da dieses RANDOM WALK-Diagramm nun alle notwendigen Informationen enthält, wäre es prinzipiell möglich, ebenfalls Zeitbetrachtungen bezüglich T_{W1} wie in 3.2 anzustellen, jedoch erwies sich die dort benutzte Methode für jenen Fall effektiver.

Es sei

- $p_F(i_1,i_2)$ die Durchlaufzustand-Wahrscheinlichkeit, mit der der Weg einer Anforderung von $\boxed{j_1,j_2}$ nach $\boxed{0,0}$ über $\boxed{i_1,i_2}$ führt (j_1,j_2 dabei implizit vorausgesetzt)
- d_i die Zahl der "p_i-Übergänge" eines Weges (p_i-Distanz) $i=1,2$.
- d_o die Zahl der Übergänge eines Weges ohne Alternative

Dann ist die Wahrscheinlichkeit eines gewissen Weges

$$P_{Weg} = 1^{d_o} \cdot p_1^{d_1} \cdot p_2^{d_2} \tag{3.69}$$

und die Durchlaufszustand-Wahrscheinlichkeit gleich der Summe der Wahrscheinlichkeiten aller Wege von $\boxed{j_1,j_2}$ zum betrachteten Durchlaufszustand.

Wenn eine Anforderung den Übergang von $\boxed{1,x}$ nach $\boxed{0,x+1}$ benutzt, trifft sie genau x Anforderungen bei Ihrer Ankunft in Stufe 2 an. Deshalb genügt es, die Durchlaufszustand-Wahrscheinlichkeiten

$$p_F(1,x) \quad (x = 0,\ldots,j_1+j_2-1)$$

zu berechnen, aus denen direkt die bedingten Antreffwahrscheinlichkeiten bei einem bestimmten Startmuster $\{x_1,x_2\}$ erhalten werden:

$$P(X_{A2}=x \mid \{x_1,x_2\}) \overset{def}{=} p_2(x \mid x_1,x_2)$$

Schließlich ergibt eine gewichtete Summation über x_2 die gesuchten Größen $p_2(x \mid x_1)$.

3.4.2 Durchlaufszustand-Wahrscheinlichkeiten

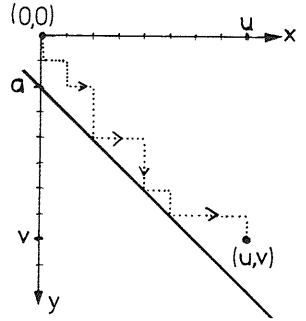
Für einen festen Start-Durchlaufszustand $\boxed{j_1,j_2}$ hängt die Wahrscheinlichkeit eines Weges zu einem festen Zustand $\boxed{1,x}$ von seiner Zahl d_o von Zuständen mit Freizeiten für Stufe 2 ab.

Für $j_1-1<x\leq j_1+j_2-1$ gibt es keinen Weg, der einen solchen Freizeitzustand berührt. Deshalb haben alle diese Wege die gleiche Wahrscheinlichkeit, ihre Anzahl ist einfach anzugeben.
Für $0<x\leq j_1-1$ gibt es Wege, die $d_o = 0,1,\ldots,d_{omax}$ Freizeitzustände besitzen. Bezeichnet man die Zahl der Wege, die von $\boxed{j_1,j_2}$ $(j_1>1, j_2\geq 0)$ zu einem Zustand $\boxed{1,x}$ $(x>0)$ über d_o Freizeitzustände führen, mit $k(d_o,x)$, so gilt

$$p_F(1,x) = \sum_{d_o=0}^{d_{omax}} k(d_o,x)\cdot p_1^{d_1}\cdot p_2^{d_2} \qquad (3.70)$$

mit $d_1=j_1-1-d_o$, $d_2=j_1-1+j_2-x$, $d_{omax}=j_1-x$.

Um $k(d_o,x)$ zu bestimmen, wird folgende Definition getroffen (vgl. Bild 3.12):



Es sei $\mathfrak{V}(d_o,u,v,a)$ die Zahl der Wege von (0,0) nach (u,v), die die Gerade $y=x+a$ genau d_o mal berühren aber nicht kreuzen.
$(a\geq 0, u>0, 0\leq v<u+a)$.

Bild 3.12 Gerichteter Weg

SATZ 3.5:

$$\mathfrak{V}(d_o,u,v,a) = \begin{cases} \binom{u+v}{u} - \binom{u+v}{u+a} & \text{für } d_o = 0 \qquad (3.71a) \\ \binom{u+v-d_o}{u+a-1} - \binom{u+v-d_o}{u+a} & \text{für } 0<d_o\leq v-a+1 \qquad (3.71b) \end{cases}$$

BEWEIS: Für $a>0, u>0, 0<v<u+a$ ist (3.71a) identisch mit "Lemma 2" von MILCH und WAGGONER |108| und kann, auch für $a\geq 0, u\geq 0, 0\leq v\leq u+a$, mit Hilfe des sog. Reflektionsprinzips (vgl.z.B. |9|) bewiesen werden. (3.17b) ist unter gleichen Voraussetzungen Teil des Satzes (Korollar 2) in |108| und wurde dort durch das sog. Teleskopprinzip bewiesen. Wegen $\mathfrak{V}(d_o,u,v,0) = \mathfrak{V}(d_o-1,u-1,v,1)$ für $d_o>0$ ist auch $a=0$ zugelassen. Dann ist die Gültigkeit für $v=0$ offensichtlich.

Geht man zurück in das RANDOM WALK-Diagramm (S.98) mit

$$k(d_o,x) = \mathfrak{V}(d_o, j_1-1, j_1-1+j_2-x, j_2) \qquad (3.72)$$

so kann nach einigen Zwischenrechnungen eine explizite Formel für die Durchlaufszustand-Wahrscheinlichkeiten $p_F(1,x)$ angegeben werden:

$$p_F(1,x) = p_2^{j_1-1+j_2-x}\left\{ p_1^{j_1-1}\cdot\binom{2j_1-2+j_2-x}{j_1-1} + p_1^x\cdot\sum_{i=0}^{j_1-1-x} p_1^{i-1}\left[\binom{j_1-2+j_2+i}{j_1-2+j_2} - p_1\cdot\binom{j_1-1+j_2+i}{j_1-1+j_2}\right]\right\} \qquad (3.73)$$

Für $x>j_1-1$ ist die Summe =0 zu setzen. $j_1\geq 1, j_2\geq 0, 0<x\leq j_1-1+j_2$

3.4.3 Bedingte Antreffwahrscheinlichkeiten

Startet eine Anforderung ihren Durchlauf mit dem Durchlaufszustand $\boxed{j_1,j_2}$, so waren bei ihrer Ankunft $x_1=j_1-1$ bzw. $x_2=j_2$ Anforderungen schon im System. Deshalb ergibt sich mit

$$p_2(x|x_1,x_2) = p_F(1,x)\cdot p_1 \qquad \text{für } x>0 \qquad (3.74)$$

aus (3.70)-(3.73)

$$p_2(x|x_1,x_2) = p_2^{x_1+x_2-x}\left\{ p_1^{x_1+1}\cdot\binom{2x_1+x_2-x}{x_1} + p_1^x\sum_{i=0}^{x_1-x} p_1^i\left[\binom{x_1+x_2-1+i}{x_1+x_2-1} - p_1\cdot\binom{x_1+x_2+i}{x_1+x_2}\right]\right\} \qquad (3.75)$$

Mit

$$p_2(0|x_1,x_2) = \begin{cases} 1 & \text{für } x_1+x_2=0 \\ p_2(1|x_1,x_2)\cdot\dfrac{p_2}{p_1} & \text{für } x_1+x_2>0 \end{cases}$$

und $\binom{-1}{-1} \overset{def}{=} 1$ kann gezeigt werden, daß (3.75) allgemein gilt für $x_1,x_2\geq 0, 0\leq x\leq x_1+x_2$.

Im speziellen Falle $\varepsilon_1=\varepsilon_2$ ist es möglich, im RANDOM WALK-Diagramm die reflektierende Grenze zu vermeiden, indem man das Diagramm durch "Spiegelung" an ihr zu einem Rechteck ergänzt (hier nicht näher beschrieben).

Dann erhält man separat für diesen Fall $\varepsilon_1=\varepsilon_2$

$$p_2(x|x_1,x_2) = \left(\frac{1}{2}\right)^{2x_1+x_2-x+1}\cdot\left[\binom{2x_1+x_2-x}{x_1}+\binom{2x_1+x_2-x}{x_1+x_2}\right], \quad x_1,x_2\geq 0, 0\leq x\leq x_1+x_2 \qquad (3.76)$$

ein Ergebnis, dessen Übereinstimmung mit (3.75) durch vollständige Induktion bewiesen werden kann.
Die bedingten Antreffwahrscheinlichkeiten für ein bestimmtes Startmuster $\{x_1,x_2\}$ können auch interpretiert werden als Zustandswahrscheinlichkeiten einer Einzelstufe M|M|1 mit Ankunftsrate ε_1 und Servicerate ε_2 (> oder $\leq \varepsilon_1$!) bei Ankunft der Anforderung Nr. x_1+1, wenn bei Beginn des Ankunftsprozesses genau x_2 Anforderungen im System waren. Dies entspricht der Untersuchung eines Zeitverhaltens, wobei "Zeit" durch die Ordnungsnummer der ankommenden Anforderung repräsentiert wird.
Diese Fragestellung wurde von TAKÁCS |20| behandelt, der mit Hilfe der Theorie der homogenen Markoff-Kette einen (umfangreichen) Ausdruck für die 2-seitige erzeugende Funktion dieser Mehrschritt-Übergangswahrscheinlichkeiten herleitete. Für einen relativ einfachen Spezialfall dieser Funktion wurde dort ein explizites Ergebnis angegeben, das, auf dieses Problem übertragen, lautet:

$$p_2(x|x_1,0) = \left(\frac{p_1}{p_2}\right)^{x}\cdot\sum_{i=x}^{x_1}\frac{i}{2x_1-i}\cdot\binom{2x_1-i}{x_1}\cdot p_1^{x_1-i}\,p_2^{x_1} \qquad (3.77)$$

Die Übereinstimmung von (3.77) und (3.75) mit $x_2=0$ kann durch vollständige Induktion gezeigt werden.

Durch gewichtete Summation über x_2 erhält man nun die gewünschten bedingten Antreffwahrscheinlichkeiten

$$p_2(x|x_1) = \sum_{x_2=0}^{\infty} P(X_{2A1}=x_2 \mid X_{A1}=x_1)\cdot p_2(x|x_1,x_2) \qquad (3.78)$$

Dabei ist X_{2A1} die zufällige Zahl von Anforderungen in Stufe 2 bei Ankunft einer Anforderung in Stufe 1.

Somit ergibt sich als allgemeine Formel

$$p_2(x|x_1) = (1-A_2)\cdot p_2^{x_1-x}\cdot\sum_{j=\max(0,x-x_1)}^{\infty}(A_2p_2)^j\cdot\left[p_1^{x_1+1}\cdot\binom{2x_1+j-x}{x_1}+ p_1^{x}\sum_{i=0}^{x_1-x}p_1^{i}\left\{\binom{x_1+j-1+i}{x_1+j-1}-p_1\cdot\binom{x_1+j+i}{x_1+j}\right\}\right], \quad x,x_1\geq 0 \qquad (3.79)$$

wobei die Summe über i für $x>x_1$ =0 zu setzen ist.
In Bild 3.13 sind diese Antreffwahrscheinlichkeiten aufgetragen und mit den absoluten Zustandswahrscheinlichkeiten $p_2(x)$ verglichen.

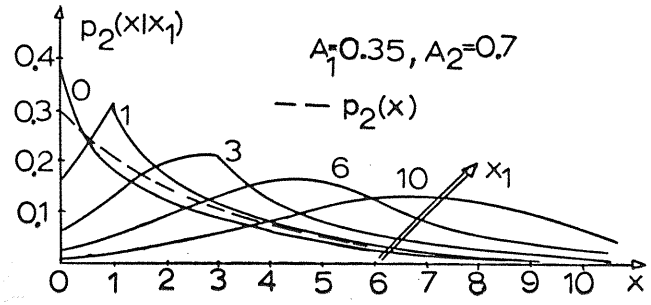


Bild 3.13 Bedingte Antreffwahrscheinlichkeiten

In diesem Beispiel, wo die 2.Stufe langsamer ist als die 1.Stufe ($\varepsilon_2<\varepsilon_1$), ist deutlich zu sehen, wie momentane Schwankungen (Verkehrsspitzen) in der 1.Stufe später in Stufe 2 fortgesetzt werden (obwohl die Durchlaufszeiten unabhängig voneinander sind!).

Es ist klar, daß für $\varepsilon_2>\varepsilon_1$ der Einfluß von x_1 auf $p_2(x|x_1)$ geringer ist und daß $p_2(x|x_1)$ für $x_1\to\infty$ gegen die absoluten Zustandswahrscheinlichkeiten einer M|M|1-Stufe mit Ankunftsrate ε_1 und Servicerate ε_2 gehen.
Für $x>x_1$ ist keine Freizeit für Stufe 2 möglich, (3.79) kann deshalb vereinfacht werden zu

$$p_2(x|x_1) = (1-A_2)\cdot A_2^{x+1}\cdot\left(\frac{1}{A_1+A_2-A_1A_2}\right)^{x_1+1}, \quad x>x_1\geq 0 \qquad (3.80)$$

was für $x_1=0$ identisch ist mit ${}_op_2^a(x)$ gemäß (3.22).
(3.79) zeigt, daß z.B. die Wartewahrscheinlichkeit in Stufe 2 umso höher ist, je größer die Zahl x_1 der angetroffenen Anforderungen in Stufe 1 war. Dies bedeutet, daß hier keine "begrenzte Abhängigkeit" gilt wie bei den Wartezeiten.

3.4.4 Weitere Schicksalsgrößen

Die Zahl der in Stufe 2 angetroffenen Anforderungen ist wegen der negativ-exponentiellen VF in Stufe 2 und der FIFO Abfertigungsdisziplin ausreichend für die Bestimmung des weiteren Schicksals in Stufe 2,wie die bedingten VF der Warte- und Durchlaufszeiten.Zum Beispiel können die bedingten mittleren Warte- und Durchlaufszeiten direkt aus dem bedingten Erwartungswert

$$E(x|x_1) \overset{def}{=} E(\, X_{A2} \,|\, X_{A1} = x_1) = \sum_{x=0}^{\infty} x \cdot p_2(x|x_1) \qquad (3.81)$$

abgeleitet werden,der in Bild 3.14a,b dargestellt ist.

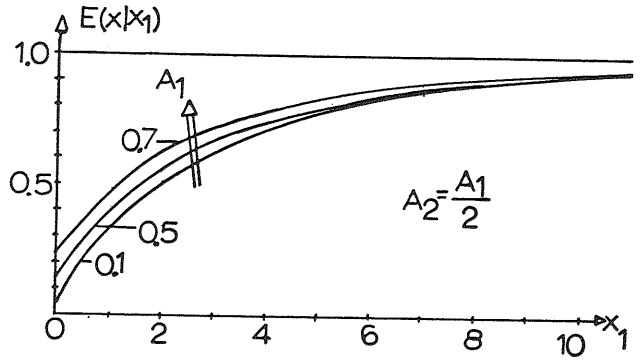


Bild 3.14a Bedingte Erwartungswerte in Stufe 2 ($\varepsilon_2 > \varepsilon_1$)

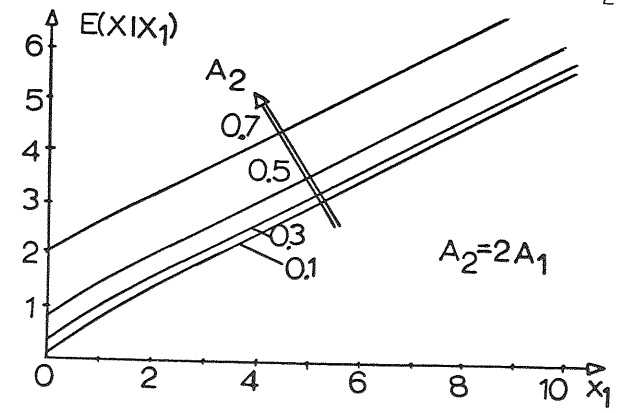


Bild 3.14b Bedingte Erwartungswerte in Stufe 2 ($\varepsilon_2 < \varepsilon_1$)

Durch diese Kurven wird folgendes prinzipielle Verhalten (das durch plausible getrennte Betrachtungen erhalten werden kann) bestätigt.
Falls $\varepsilon_2 > \varepsilon_1$,gilt entsprechend einem quasistationären Verhalten

$$\lim_{x_1 \to \infty} E(x|x_1) = \frac{A_2^*}{1-A_2^*} \text{ mit } A_2^* = \frac{\varepsilon_1}{\varepsilon_2} \qquad (3.82)$$

wogegen für $\varepsilon_2 < \varepsilon_1$ dieser Erwartungswert mit x_1 nahezu linear ansteigt mit

$$\frac{E(x|x_1+\Delta x_1) - E(x|x_1)}{\Delta x_1} \approx 1 - \frac{\varepsilon_2}{\varepsilon_1} \qquad (3.83)$$

da die Wahrscheinlichkeit für eine leere 2.Stufe $\to 0$ geht für genügend großes ε_1 und x_1.

Betrachtet man nun Gesamtschicksale wie in 3.3,so können außer einiger bereits dort abgeleiteten Größen die Gesamtzahl der Vorgänger-Anforderungen in Stufe 1 und 2

$$P(\, X_{A1} + X_{A2} = x) = \sum_{j=0}^{x} P(\, X_{A1} = x-j) \cdot p_2(j|x-j) \qquad (3.84)$$

bestimmt werden,die angibt,aus wievielen Wartephasen die Gesamtwartezeit einer Anforderung besteht.

4 BERECHNUNG 2-STUFIGER SYSTEME OHNE BLOCKIERUNG BEI NICHT-REKURRENTEN VERKEHREN ZWISCHEN DEN STUFEN

In diesem Kapitel wird für das 2-stufige Single-Server Wartesystem ohne Blockierung

$$M \xrightarrow{\infty} VF_1 \mid 1 \xrightarrow{\infty} VF_2 \mid 1 \qquad (VF_1, VF_2 \text{ beliebig hypoexp.})$$

eine Approximationsformel für die mittleren Wartezeiten aller Anforderungen in Stufe 2 hergeleitet.Da Stufe 1 für sich exakt berechnet werden kann,sind deshalb auch die Mittelwerte der Gesamtdurchlaufs- und Wartezeiten angebbar.
Es werden in diesem Zusammenhang nur sog. hypoexponentielle VF betrachtet und eine verkehrs- und varianzabhängige Interpolation zwischen Ergebnissen von bekannten exakten oder neuen approximativen (Teil-) Ergebnissen vorgenommen.Die entwickelte Näherung wird mit Simulationsergebnissen verglichen,außerdem mit 2 naheliegenden einfacheren (Vergleichs-) Näherungen.

4.1 Einführung

Bei seriellen Wartesystemen mit unendlich großen Zwischenspeichern tritt nirgendwo Blockierung von Bedienungseinheiten (BE) auf.Trotzdem kann bei diesen Systemen der Ankunftsprozeß in einer Folgestufe nicht allein durch einen unabhängigen Ankunftsprozeß mit entsprechender VF der Ankunftsabstände beschrieben werden,da (vgl. 2.2.1) im allgemeinen aufeinanderfolgende Abgangsabstände von Stufe i (= Ankunftsabstände in Stufe i+1) voneinander abhängig sind,d.h. kein rekurrenter Eingangsprozeß mehr vorliegt.Demzufolge können die Standardmethoden für die Untersuchung von einstufigen Wartesystemen (vgl. 1.2.3) nicht oder zumindest nur näherungsweise verwendet werden.
Selbst bei (angenommener) Unabhängigkeit aufeinanderfolgender Ankunftsabstände in einer Folgestufe ist für ihre Berechnung die Bestimmung der Verkehrsgrößen in einem GI|G|1-System notwendig.
In Abschnitt 2.2 wurde eine ausführliche Übersicht über den in diesem Kapitel behandelten engeren Themenkreis gegeben und insbesondere in 2.2.3 die (gescheiterten) Versuche von Autoren geschildert,die Wartezeiten in Stufe 2 exakt zu bestimmen,

selbst unter der Annahme einer negativ-exponentiellen VF der Bedienungszeiten in Stufe 2.
Diese Darlegungen stellen für das Verständnis dises Kapitels 4 eine notwendige Voraussetzung dar.Deshalb zeigt Tabelle 4.1 hierzu nur kurz die in der Literatur behandelten entsprechenden 2-stufigen Systeme.

Veröffentlichung	VF_1 - VF_2	Ergebnisse f. Stufe 2
BURKE |30|	M - G	Berechnung nach M|G|1 (exakt)
GHOSAL |66|	E_2 - M	Wartezeit-VF (nicht exakt,Abhängigkeiten unberücksichtigt)
SUZUKI |102|	G - M	
LOYNES |81|	E_2 - M	Wartezeit-VF ist nicht negativ-exponentiell
FRIEDMAN |65|	D - D	$E(T_{W2})$ exakt aus Reduktionstheorie
Diese Arbeit	bel.hypoexp.	$E(T_{W2})$,approximativ

Tabelle 4.1 Behandelte 2-stufige Single-Server Systeme ohne Blockierung

In diesem Kapitel werden beliebige hypoexponentielle Verteilungsfunktionen (VF) in beiden Stufen zugelassen und diese näherungsweise durch ihre ersten beiden Momente (Mittelwert und Varianz bzw.Varianzkoeffizient) beschrieben.Dies ist z.B. auch bei Überlaufsystemen üblich (vgl.z.B. |14|) und liefert für die Praxis genügend genaue Ergebnisse.
Da im Anhang A2 generell auf den Einfluß des 3. und höherer Momente bei den in dieser Arbeit untersuchten Systemen eingegangen wird,soll auf diesen hier nicht eingegangen werden (vgl. dort).

4.2 Einfache Möglichkeiten der Approximation

Es werden hier zunächst 2 einfache Näherungsmöglichkeiten geschildert.die sowohl als Vergleich zu den Ergebnissen der speziellen Näherungsmethode in 4.3 dienen,als auch zur Motivation des dort verwendeten Näherungsprinzips.

4.2.1 Annahme eines Poisson-Ankunftsprozesses (Näherung P)

Als einfachste und zugleich gröbste Näherung kann die pauschale Annahme eines Poisson-Ankunftsprozesses für Stufe 2 (Näherung P) gelten. Diese näherungsweise Beschreibung von Stufe 2 als $M|VF_2|1$-System ist umso genauer, je mehr die VF der Bedienungszeiten in Stufe 1 VF_1 zu einer negativ-exponentiellen VF (Varianzkoeffizient $C_{H1}=1$) neigt. Die Güte dieser Näherung hängt bei gegebenen VF-Typen der Bedienungszeiten in beiden Stufen auch vom Verhältnis h_1/h_2 ihrer Mittelwerte ab: je kleiner h_1/h_2, desto mehr "schlägt der Poissonprozeß am Eingang von Stufe 1 auf ihren Ausgang durch".
Damit ergibt sich für den Erwartungswert der Wartezeit in Stufe 2 - gemäß den Formeln für das Wartesystem M|G|1 nach POLLACZEK und KHINTCHINE, vgl. z.B. |19| -

$$E(T_{W2}) = \frac{(1 + C_{H2}^2)\cdot A_2}{2(1 - A_2)}\cdot h_2 \qquad \text{(Näherung P)} \qquad (4.1)$$

$$\text{mit } A_2=\lambda\cdot h_2$$

Weil bei hypoexponentiell verteilten Bedienungszeiten der Ausgangsprozeß einen Glättungseffekt gegenüber dem Poisson-Eingang aufweist (bzw. bei hyperexponentiellen VF einen Spitzigkeitseffekt, vgl. 2.2.1), ist zu vermuten, daß die mit dieser Näherung erhaltenen mittleren Wartezeiten in Stufe 2 bei dem hier betrachteten Fall $C_{H1} \leqslant 1$ überschätzt (im Falle $C_{H1} > 1$ unterschätzt) werden, was sich in den Bildern 4.3, 4.5-4.7 bestätigt (S.115 ff).

Bei diesem Näherungsverfahren P würde sich als Konsequenz der Poisson-Annahme für alle VF_1-VF_2-Kombinationen eine von der Reihenfolge der beiden Stufen unabhängige mittlere Gesamtwarte- und Durchlaufszeit ergeben; dies ist zwar für D-D und M-M gegeben, jedoch im Regelfalle nicht erfüllt, da bei hypoexponentiellen VF in Stufe 1 der Poisson-Ankunftsprozeß vor Stufe 1 als geglätteter Ausgangsprozeß am Ausgang dieser Stufe erscheint und demzufolge die mittlere Wartezeit aller Anforderungen in Stufe 2 kleiner ist als bei Poisson-Ankünften.
Dieser Glättungseffekt ist bekannt und wurde auch von PACK |91| anhand eines 2-stufigen Modells für die Übertragung von Daten auf einer Multiplex-Übertragungsleitung zu einem Rechner (vgl. 1.4.2) beschrieben. Die von ihm durch Simulation gewonnenen Werte für die Reduktion der Wartezeiten in Stufe 2 eines Systems mit D-M durch den Glättungseffekt konnten bestätigt werden.

4.2.2 Annahme eines allgemeinen rekurrenten Ankunftsprozesses (Näherung R)

Der Glättungseffekt beim Zwischenverkehr kann berücksichtigt werden, indem zunächst die Varianz $\sigma_{A2}^2 = C_{A2}^2/\lambda^2$ des Eingangsprozesses in Stufe 2 nach der Formel von MAKINO für den Ausgangsprozeß eines Systems M|G|1 (vgl. 2.2.1) nach (2.1) bestimmt wird. Danach beträgt der Varianzkoeffizient des Eingangsprozesses in Stufe 2

$$C_{A2} = \sqrt{1 - A_1^2\cdot(1 - C_{H1}^2)} \qquad (4.2)$$

Zur Berechnung der 2.Stufe als Einzelstufe mit einem Varianzkoeffizient C_{A2} des Eingangsprozesses und Bedienungszeit-Varianzkoeffizient C_{H2} (Näherung R) muß jedoch eine Unabhängigkeit aufeinanderfolgender Ankunftsabstände angenommen werden (rekurrenter Prozeß).
Für dieses allgemeine einstufige reine Wartesystem mit 1 BE und beliebigen VF für den Ankunfts- und Bedienungsprozeß (GI|G|1) sind in der Literatur verschiedene Methoden zur Berechnung der Verkehrsgrößen bekannt. Eine wichtige Methode ist z.B. die Lösung der Integralgleichung nach LINDLEY |40|, die jedoch nicht allgemein, d.h. für beliebige VF angegebeben werden kann.
Deshalb wurde in |39| eine einfache Approximationsformel für die mittlere Wartezeit aller Anforderungen in einem GI|G|1-System entwickelt, die nur die ersten beiden Momente der VF des Ankunfts- und Bedienungsprozesses berücksichtigt:

$$E(T_W)_{GI|G|1} = \begin{cases} \dfrac{A}{(1-A)}\cdot\dfrac{(C_A^2 + C_H^2)}{2}\cdot e^{-\left\{\frac{2(1-A)}{3A}\cdot\frac{(1-C_A^2)^2}{(C_A^2+C_H^2)}\right\}}\cdot h & C_A \leqslant 1 \quad (4.3) \\ \dfrac{A}{(1-A)}\cdot\dfrac{(C_A^2 + C_H^2)}{2}\cdot e^{-\left\{(1-A)\cdot\frac{(C_A^2-1)}{(C_A^2+4C_H^2)}\right\}}\cdot h & C_A \geqslant 1 \quad (4.4) \end{cases}$$

Diese Formeln enthalten verschiedene in der Literatur bekannte (exakte und approximative) Ergebnisse und wurden anhand vieler Systeme mit D, E_k, M und H_2 VF mit Simulationsprogrammen getestet.

Setzt man nun in (4.3) für C_A den Varianzkoeffizienten C_{A2} nach (4.2) ein, erhält man die gewünschte einfache Vergleichsnäherung R.

Die hiermit erhaltenen Ergebnisse (vgl. Bilder 4.3, 4.5-7 S.115 ff.) sind durchweg besser als bei Näherung P.
Jedoch sind die Ergebnisse zum Teil noch sehr unbefriedigend, was sich insbesondere bei Bedienungszeit-VF mit kleinen Varianzkoeffizienten bemerkbar macht. Dies ist besonders aus dem (hier nicht dargestellten) Grenzfall konstanter Bedienungszeiten in beiden Stufen zu ersehen, wo im Falle $h_2<h_1$ nie eine Anforderung in Stufe 2 warten muß, mit der Vergleichsnäherung R sich aber für $0 < A_1 < 1$ mittlere Wartezeiten >0 ergeben. Dies ist darauf zurückzuführen, daß in diesen Fällen der Ankunftsabstand in Stufe 2 stets größer ist als die Bedienungszeit ($T_{A2}>T_{H2}$), dies aber bei einer 2-Momentendarstellung des Ankunftsprozesses nicht vollständig berücksichtigt sein kann.
Wie aus dem Verlauf von Näherung R in Bild 4.5 unten ersichtlich, ist auch bei D-E_2 mit $h_2<h_1$ noch etwas von dieser Tendenz zu spüren. Diese Maxima, die nur bei $h_2<h_1$ beobachtet wurden, können natürlich nur deshalb auftreten, weil sich mit steigender Ankunftsrate auch der Typ der VF des Ankunftsprozesses in Stufe 2 ändert.
Im Grenzfall $A_1 \to 1$ kann Stufe 2 exakt als eine Einzelstufe $VF_1|VF_2|1$ beschrieben werden.
Würde man in Stufe 2 die Ankunftsabstände an sich durch ihre VF

$$P(T_{A2}>t) = W_1 \cdot P(T_{H1}>t) + (1-W_1)\cdot P(T_{A1}+T_{H1}>t)$$

exakt berücksichtigen -also nur die Abhängigkeiten zwischen ihnen vernachlässigen- so würden sich (beispielsweise durch Simulation) mittlere Wartezeiten ergeben, die je nach System bis zu ca 35% (bei D-D) unterhalb der exakten Werte liegen.
Jedoch kann gesagt werden, daß die Unterschätzung der mittleren Wartezeiten durch Vernachlässigung der Abhängigkeiten aufeinanderfolgender Ankunftsabstände in Stufe 2 von h_1/h_2 abhängt und umso geringer ist, je besser die Bedienungszeiten in Stufe 1 durch eine negativ-exponentielle VF zu beschreiben wären.

Diese gravierenden Nachteile der hier nur zum Vergleich geschilderten Näherungen P und R führten zu dem gewählten Verfahren, das nun beschrieben wird.

4.3 Genaueres Näherungsverfahren (Näherung Ip)

4.3.1 Prinzip der Näherung

Wie aus der Zusammenstellung der in der Literatur untersuchten 2-stufigen Single-Server Systeme ohne Blockierung

$$M \xrightarrow{\infty} VF_1|1 \xrightarrow{\infty} VF_2|1$$

in 2.2 und 4.1 hervorgeht, existieren exakte Ergebnisse für die mittleren Wartezeiten aller Anforderungen in Stufe 2 $E(T_{W2})$ nur für folgende VF_1-VF_2-Kombinationen:

- M-G Da der Ausgangsprozeß von Stufe 1 ein Poissonprozeß ist, kann die 2.Stufe exakt wie eine M|G|1-Stufe berechnet werden (nach (4.3) für Näherungsverfahren P, das für diesen Fall exakt ist).
- D-D Nach der Reduktionstheorie von FRIEDMAN u. AVI-ITZHAK ist die gesamte Verzögerungszeit T_{WB} (hier = Wartezeit) einer Anforderung in diesem Falle gleichgroß wie die Wartezeit im einstufigen System M|D|1 mit $h = \max(h_1,h_2)$. Also beträgt die mittlere Wartezeit aller Anforderungen in Stufe 2

$$E(T_{W2})_{D-D} = E(T_{WB}) - E(T_{W1})$$

$$= \begin{cases} \frac{A_2}{2(1-A_2)}\cdot h_2 - \frac{A_1}{2(1-A_1)}\cdot h_1 & h_1 \leq h_2 \\ 0 & h_1 \geq h_2 \end{cases} \qquad (4.5)$$

In Bild 4.1 sind die hier zunächst untersuchten VF_1-VF_2- Kombinationen in Matrixform entsprechend ihrer Varianzkoeffizienten dargestellt und bei Systemen mit bekannten exakten (expliziten) Ergebnissen eingerahmt.

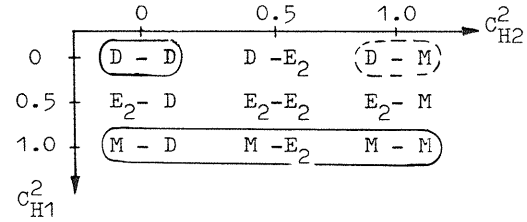


Bild 4.1 Hypoexponentielle VF_1-VF_2-Kombinationen der Bedienungszeiten (◯ explizit angebbar)

Gelingt es, für die Kombination (D-M) eine Formel für die mittlere Wartezeit in Stufe 2 herzuleiten, so kann für alle Systeme mit $0 \leqslant C_{H1} \leqslant 1$ und $0 \leqslant C_{H2} \leqslant 1$ eine varianzabhängige Interpolation zwischen diesen Ergebnissen erfolgen (Näherung Ip).

4.3.2 Approximation für Bedienungszeit-VF-Kombination D-M

Der Grundgedanke der approximativen Lösung dieses Teilproblems der Bestimmung der mittleren Wartezeit in Stufe 2 besteht in einer einer vom Angebot A_1 abhängigen Interpolation zwischen den Wartezeiten nach folgenden Berechnungsprinzipien "D" und "O".

- D: Die mittleren Wartezeiten in Stufe 2 werden berechnet, als ob bei gleicher Durchsatzrate die Ankunftsabstände in Stufe 2 konstant wären (Stufe 2 als D|M|1-Stufe).
- O: Die mittleren Wartezeiten werden aus einer noch herzuleitenden Formel für "kleine" Angebote in Stufe 1 für das betreffende Angebot A_1 berechnet.

Das Berechnungsprinzip O wird umso genauer sein, je kleiner A_1 ist, während Prinzip D bei $h_1 \geqslant h_2$ für $A_1 \to 1$ exakt ist, der zugehörige Gewichtungsfaktor also gleich 1 sein muß, wogegen bei $h_1 < h_2$ -wo selbst bei Grenzbelastung des Systems die BE der Stufe 1 nicht immer belegt ist- Prinzip D eine mehr oder weniger grobe Näherung darstellt. Diese Teilnäherung ist aber dann von untergeordneter Bedeutung, wenn sie nur mit einem geringen Gewicht in die gesamte Näherung des Teilproblems eingeht.

- Bei der Annahme konstanter Ankunftsabstände in Stufe 2 (Näherung D), ergibt sich aus (4.3) approximativ

$$E(T_{W2})_D = E(T_{W2})_{D|M|1} = \frac{A_2}{2(1-A_2)} \cdot e^{-\frac{2(1-A_2)}{3A_2}} \cdot h_2 \qquad (4.6)$$

- Bei der Herleitung einer Näherungsformel für Prinzip O wurde zunächst davon ausgegangen, daß bei kleinen Angeboten A_1 und damit kleiner Wartewahrscheinlichkeit $W_1 = A_1$ in Stufe 1 der Ankunftsprozeß in Stufe 2 approximativ als Poissonprozeß angenommen werden darf. Als Grundlage hierzu diente
  a) die Anschauung, daß dann wegen kleiner Wartewahrscheinlichkeit in Stufe 1 die mit kleiner Rate ankommenden Poisson-Ankünfte jeweils nur um eine konstante Zeit h_1 verzögert werden

  b) die Tatsache, daß der Ausgangsprozeß von $M|G|\infty$ ein Poisson-Prozeß ist (vgl. MIRASOL |43|) und bei $\lambda \to 0$ die Zahl der BE unerheblich ist.

Es zeigte sich jedoch durch Vergleich mit Simulationsergebnissen, daß diese Berechnung der 2. Stufe als M|M|1-Stufe für kleine Angebote A_1 relativ gesehen zu hohe Ergebnisse lieferte (vgl. z.B. Bild 4.3 unten). Dabei ist klar, daß für $A_1 \to 0$ $E(T_{W2})$ ebenfalls -und zwar unabhängig von Annahmen bezüglich des Ausgangsprozesses von Stufe 1- absolut gesehen gegen 0 geht. Was sich aber in obigem Falle als unbefriedigend herausstellte, war der Effekt, daß der Grenzwert

$$\lim_{\lambda \to 0} \left\{ \frac{E(T_{W2})_{D-M}}{E(T_{W2})_{M|M|1}} \right\} \overset{def}{=} F_{0\ D-M} \qquad (4.7)$$

nicht gegen 1 zu gehen scheint.

Daß ein solcher Grenzwert $\neq 1$ möglich ist, kann man z.B. für D-D, also konstanten Bedienungszeiten in Stufe 2, mit (4.5) exakt zeigen.

$$\lim_{\lambda \to 0} \left\{ \frac{E(T_{W2})_{D-D}}{E(T_{W2})_{M|D|1}} \right\} \overset{def}{=} F_{0\ D-D} = \begin{cases} 1 - \frac{h_1}{h_2} & h_1 \leqslant h_2 \\ 0 & h_1 \geqslant h_2 \end{cases} \qquad (4.8)$$

Zur Bestimmung des Verlaufes von $F_{0\ D-M}$ bot es sich an, auf einem Ergebnis von SUZUKI |102| für D-M aufzubauen, bei dem allerdings Abhängigkeiten nicht berücksichtigt wurden. Danach gilt speziell für diesen Fall

$$\left.\begin{aligned} E(T_{W2}) &= \frac{\omega}{1-\omega} \cdot h_2 \\ \text{wobei } \omega \text{ die Wurzel der Gleichung} & \\ e^{-\frac{h_1}{h_2}(1-\omega)} \cdot \frac{\lambda + \lambda \cdot \frac{h_1}{h_2}(1-\omega)}{\lambda + \frac{1}{h_2}(1-\omega)} &= \omega \end{aligned}\right\} \qquad (4.9)$$

darstellt. Für $\lambda \to 0$ verschwinden die nichtberücksichtigten Abhängigkeiten, aber auch ω geht gegen 0.
Setzt man nun in 1. Näherung

$$\lim_{\lambda \to 0} \frac{\omega}{\lambda} = h_0$$

und geht damit in (4.9) ein, so ergibt sich für den Grenzwert

$$h_0 = (h_1 + h_2) \cdot e^{-h_1/h_2}$$

und daraus

$$F_{0\ D-M} = \frac{h_0}{h_2} = \frac{h_1 + h_2}{h_2} \cdot e^{-h_1/h_2} \qquad (4.10)$$

Dieser Verlauf ist in Bild 4.2 gestrichelt dargestellt.

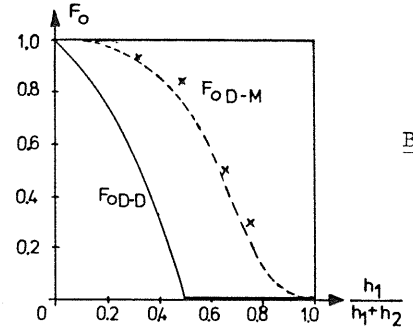


Bild 4.2 Verlauf des Grenzwerts $F_{0\,D-M}$

Die eingetragenen Kreuze stellen etwa den ermittelten Verlauf von $F_{0\,D-M}$ dar, der durch (4.10) zufriedenstellend beschrieben wird. Er wurde gewonnen durch Simulation der Systeme für $A_{max} = \max(A_1, A_2) = 0.2 \ldots 0.8$ bzw 1.0 und Extrapolation der (hier nicht dargestellten) Verbindungskurven der Simulationsergebnisse auf $A_{max} \approx 0.1$ und Vergleich mit den Ergebnissen der Näherung P an derselben Stelle.

Damit beträgt die mittlere Wartezeit in Stufe 2 für kleine Angebote A_1 (Näherung 0) approximativ

$$E(T_{W2})_0 = F_{0\,D-M} \cdot E(T_{W2})_{M|M|1} = F_{0\,D-M} \cdot \frac{A_2}{1-A_2} h_2 \qquad (4.11)$$

Untersucht man die Ergebnisse beider Berechnungsprinzipien nach (4.6) und (4.11), so ergeben sich prinzipiell die in Bild 4.3 unten zusätzlich dargestellten strichpunktierten Verläufe.
Es verbleibt die von A_1 abhängige Interpolation zwischen diesen beiden Ergebnissen:

$$E(T_{W2})_{D-M} = (1-g(A_1)) \cdot E(T_{W2})_0 + g(A_1) \cdot E(T_{W2})_D \qquad (4.12)$$
$$\text{mit } 0 < g(A_1) < 1,\ g(0) = 0,\ g(1) = 1$$

Durch Vergleich mit Simulationsergebnissen stellte sich die Notwendigkeit heraus, den Interpolationsfaktor $g(A_1)$ auch vom Verhältnis h_2/h_1 der mittleren Bedienungszeiten abhängig zu machen. Als einfach und brauchbar erwies sich

$$g(A_1) = A_1^{2\left(\frac{h_2}{h_1}+1\right)} \qquad (4.13)$$

Damit ist es nun gelungen, eine näherungsweise Berechnung der mittleren Wartezeit $E(T_{W2})_{D-M}$ für die Kombination D-M zu finden.

In Bild 4.3 sind die Ergebnisse dargestellt für $h_1/h_2 = 1$ und 2, d.h. bei $h_1+h_2 = 1$ Zeiteinheit für $h_1 = 0.5$ und $0.\overline{6}$. Im Falle $h_1 = 0.\overline{3}$ lagen alle Ergebnisse des Näherungsverfahrens innerhalb der Vertrauensintervalle der Simulation.

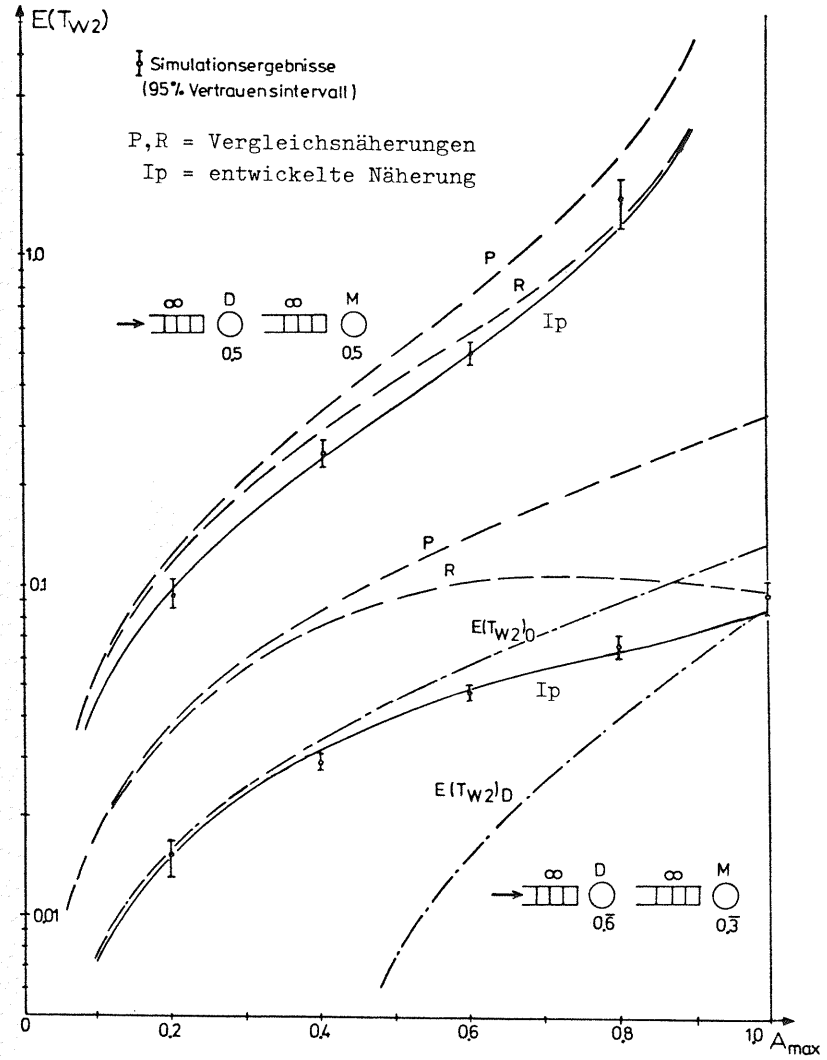


Bild 4.3 Mittlere Wartezeit $E(T_{W2})$ für D-M

4.3.3 Approximation für beliebige hypoexponentielle VF-Kombinationen

Wie bei der Beschreibung des Näherungsprinzips in 4.3.1 angedeutet, erhält man eine Näherungsformel für beliebige hypoexponentielle VF in beiden Stufen durch eine von beiden Varianzkoeffizienten abhängige Interpolation zwischen den exakten Ergebnissen für M-G, D-D und den in 4.3.2 abgeleiteten Näherungsergebnissen für D-M (vgl. Bild 4.1).
Als Interpolationsfaktoren wurden (wie bei den mittleren Wartezeiten im System M|G|1, vgl. (4.1)) **nur vom Quadrat der Varianzkoeffizienten linear abhängige Terme** in Aussicht genommen und bei zufriedenstellender Genauigkeit beibehalten.
Bild 4.4a,b zeigt 2 Simulationsreihen, wo bei fester Ankunftsrate die Varianz der VF in Stufe 1 bzw. 2 variiert wurde.

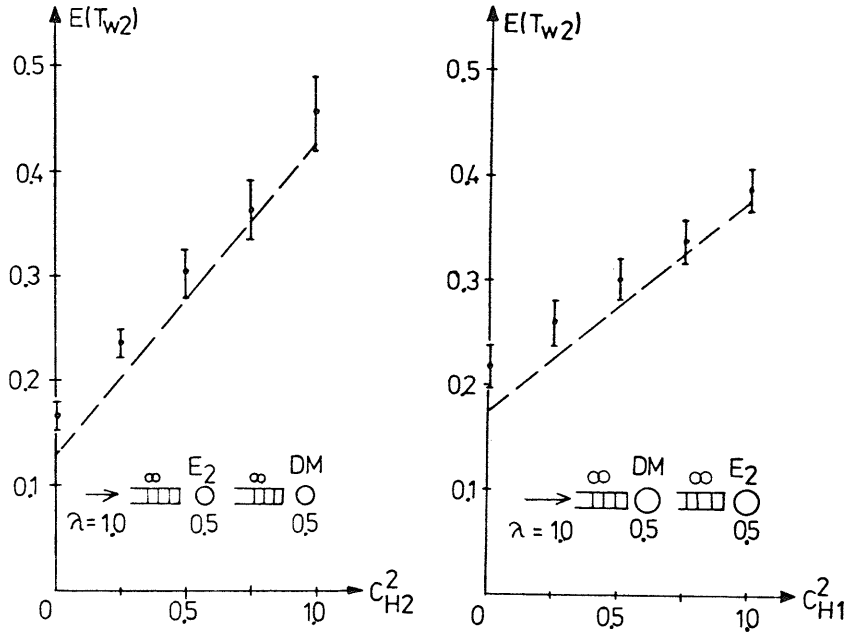


Bild 4.4a,b
a) Variation von VF_2 bei $VF_1=E_2$ ($h_1=h_2=0.5$)
b) Variation von VF_1 bei $VF_2=E_2$ ($h_1=h_2=0.5$)

Als variierte VF wurde eine hier "DM" genannte VF gewählt (Summe aus konstantem Wert + negativ-exponentieller Phase, vgl. A1.6).

Die eingezeichneten Geraden (lineare Interpolation) wurden gewonnen aus

$$E(T_{W2})_{VF1-VF2} = (1-C_{H1}^2)\cdot(1-C_{H2}^2)\cdot E(T_{W2})_{D-D} + (1-C_{H1}^2)C_{H2}^2 E(T_{W2})_{D-M} + C_{H1}^2\cdot(1-C_{H2}^2)\cdot E(T_{W2})_{M-D} + C_{H1}^2 C_{H2}^2 E(T_{W2})_{M-M} \quad (4.14)$$
$$= (1-C_{H1}^2)\cdot E(T_{W2})_{D-VF2} + C_{H1}^2\cdot E(T_{W2})_{M-VF2}$$

mit

$$E(T_{W2})_{D-VF2} = (1-C_{H2}^2)\cdot E(T_{W2})_{D-D} + C_{H2}^2\cdot E(T_{W2})_{D-M}$$

Setzt man die einzelnen Erwartungswerte nach (4.1), (4.5) und (4.10) -(4.13) ein, erhält man die explizite Formel

$$E(T_{W2})_{VF1-VF2} = (1-C_{H1}^2)\left\{(1-C_{H2}^2)\cdot\max\left[0, f(A_2)-f(A_1)\right] + C_{H2}^2 f(A_2)\cdot\left(2\cdot[1-g(A_1)]\cdot F_{oD-M} + g(A_1)\cdot e^{-\frac{2(1-A_2)}{3A_2}}\right)\right\} + C_{H1}^2(1+C_{H2}^2)\cdot f(A_2)$$

wobei

$$f(A_i) = \frac{A_i}{2(1-A_i)}\cdot h_i \qquad i=1,2 \quad ; \quad g(A_1) = A_1^{2\left(\frac{h_2}{h_1}+1\right)} \quad (4.15)$$

F_{oD-M} nach (4.10)

In den Bildern 4.5-4.7 sind einige weitere Bedienungszeit-VF-Kombinationen für jeweils 2 Werte der mittleren Bedienungszeiten je Stufe dargestellt; weitere Kurven siehe Anhang A2.

4.3.4 Güte und Gültigkeitsbereich der Näherung

Mit Hilfe der Simulation wurden zunächst alle 15 Systeme untersucht mit $VF_{1,2}=D, E_2, M$ und $h_1/h_2=1/2, 1, 2$, von denen keine exakte Lösung bekannt ist, bei denen also (4.15) nur eine Näherung darstellt.

- Für $h_1/h_2 \to 0$ sind die Ergebnisse für alle VF-Kombinationen exakt, als mittlere Wartezeit erhält man diejenigen für $M|VF_2|1$.
- Für $0<h_1/h_2<1/2$ kann gesagt werden, daß der maximale Fehler für den Angebotsbereich $A_{max}=0.2...0.8$ wegen der Annäherung des Ausgangsprozesses an einen Poissonprozeß kleiner ist als der Fehler bei $h_1/h_2=1/2$ und der (ungünstigsten) Kombination D-M. Dort lagen alle Näherungswerte innerhalb der Vertrauensintervalle der

Simulation, die maximal den relativen Wert von ca + 12% besaßen.

● Für $h_1/h_2=1$ war der relative Fehler bezogen auf den Mittelwert der Simulation stets $\lesssim$ 20% für alle obigen Kombinationen.

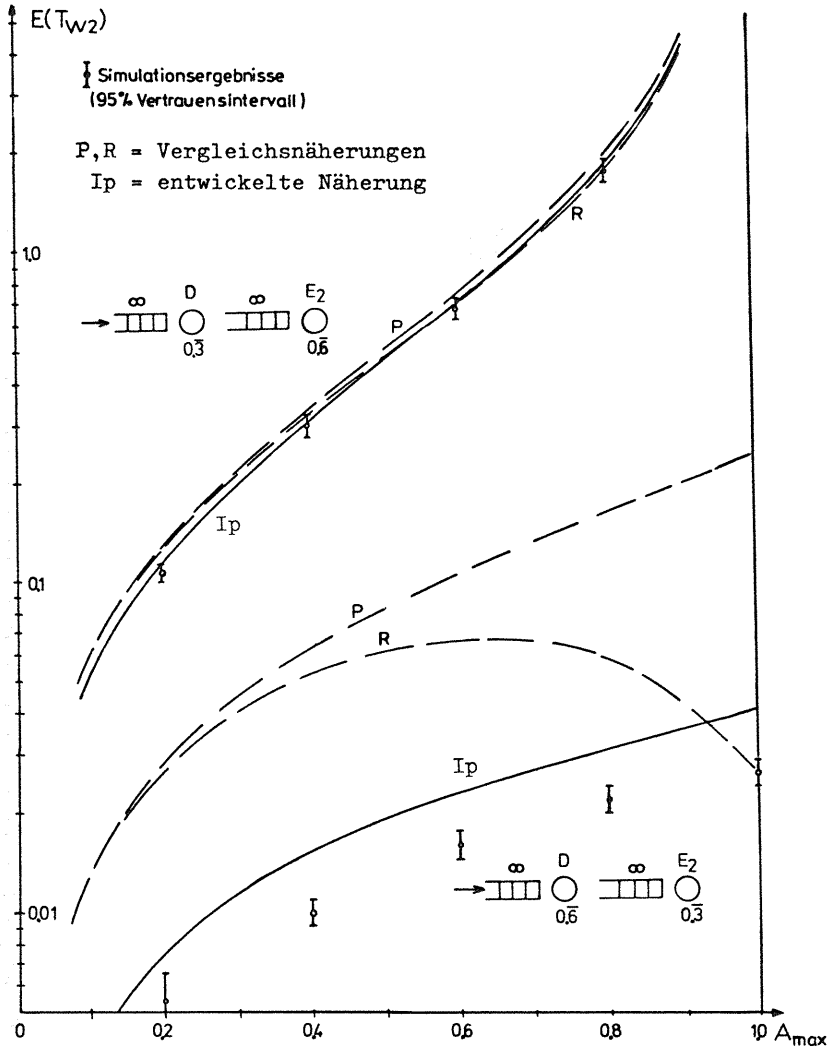


Bild 4.5 Mittlere Wartezeit $E(T_{W2})$ für $D\text{-}E_2$

● Für $h_1/h_2=2$ wiesen nur die Kombinationen $D\text{-}E_2$ und $E_2\text{-}D$ ($C^2_{H1}+C^2_{H2}=0.5$) einen größeren Fehler auf, nämlich $\lesssim$50% bzw. $\lesssim$30%. Diese Fehler sind jedoch nicht gravierend, da sie auftreten

- im Bereich $E(T_{W2})/h_2 \lesssim 1/10$ bzw. 1/6
- bei einem Verhältnis $E(T_{W2})/E(T_{W1}) \lesssim 1/20$ bzw. 1/50

und deshalb fast keinen Einfluß auf die gesamte Wartezeit $E(T_W)=E(T_{W1})+E(T_{W2})$ bzw. die gesamte Durchlaufszeit von Anforderungen haben.

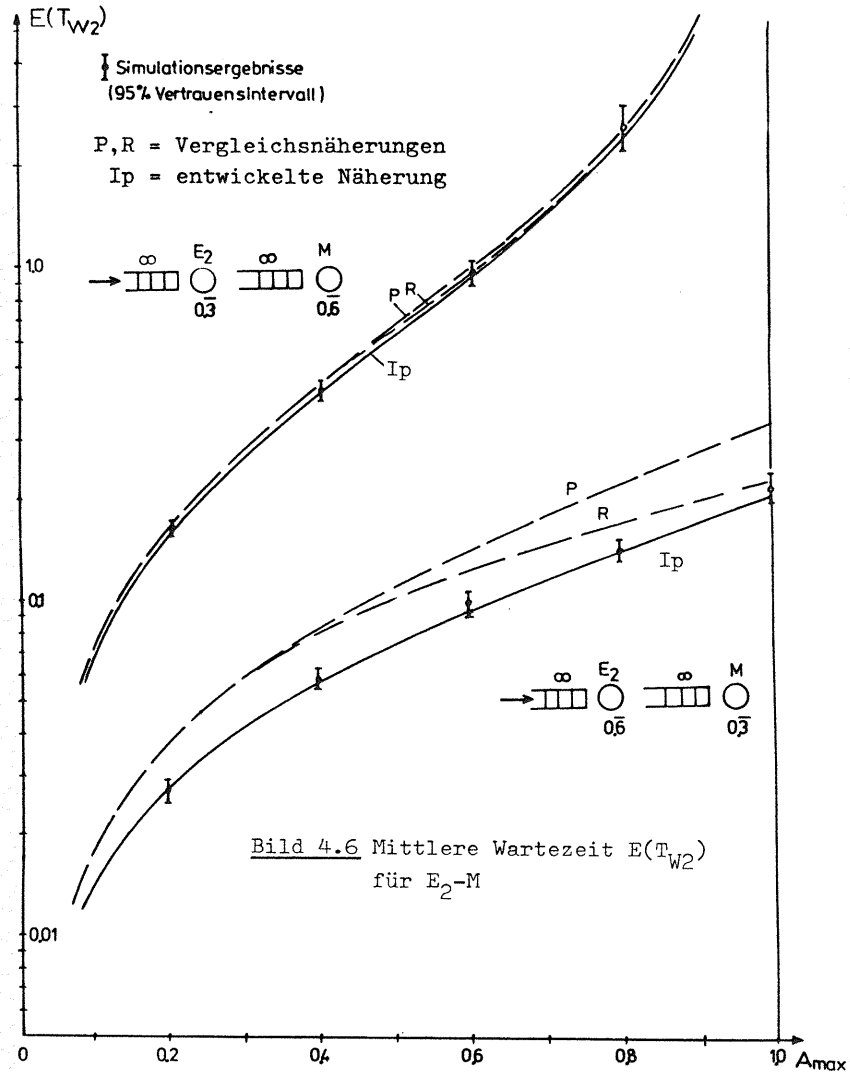


Bild 4.6 Mittlere Wartezeit $E(T_{W2})$ für $E_2\text{-}M$

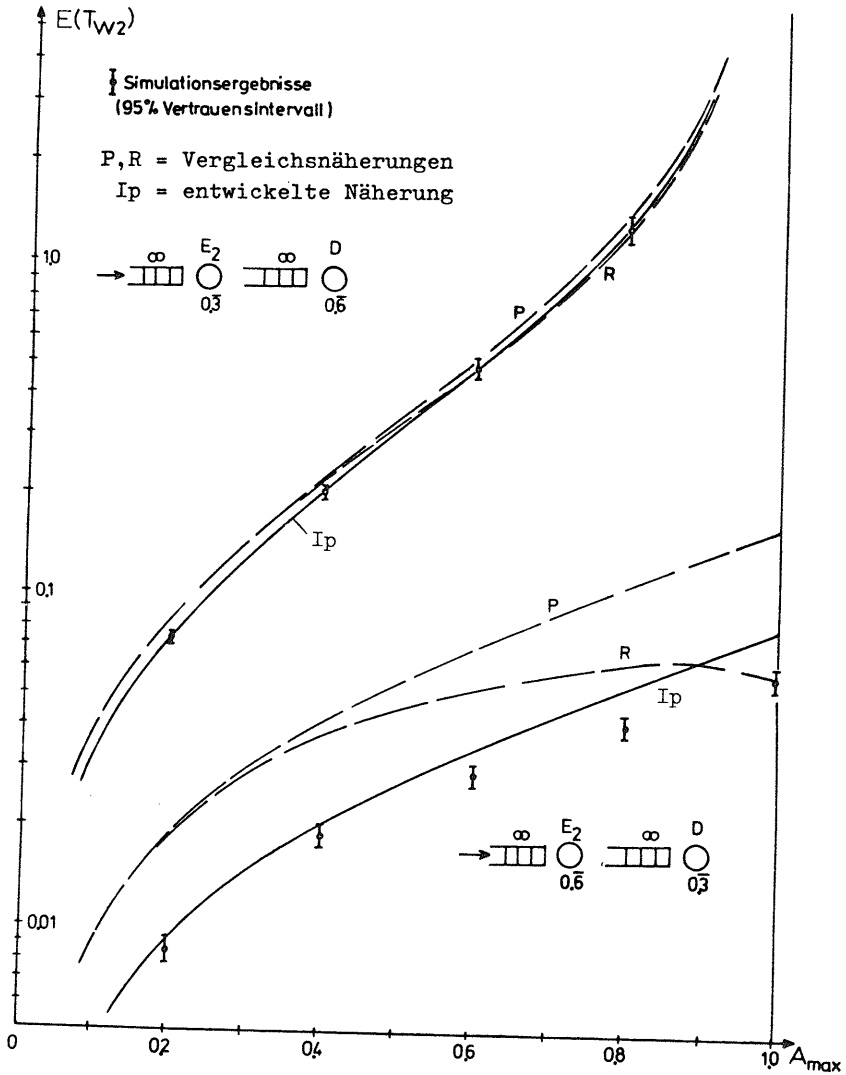


Bild 4.7 Mittlere Wartezeit $E(T_{W2})$ für E_2-D

Bezüglich der Genauigkeit des Verfahrens bei anderen VF-Typen und den Einfluß höherer Momente vgl. Anhang A2.

5 MAXIMALER DURCHSATZ BEI 2 BEDIENUNGSEINHEITEN IN SERIE MIT ENDLICH GROSSEM ZWISCHENSPEICHER

In diesem Kapitel wird eine einfache Näherungsformel für den maximalen Durchsatz ermittelt,der bei 2 seriell angeordneten Bedienungseinheiten (BE) mit endlich großem Zwischenspeicher erzielt werden kann.Die Approximation,die bekannte exakte Ergebnisse enthält,ist anwendbar auf beliebige VF der Bedienungszeiten in beiden BE,jedoch werden nur deren ersten beiden Momente (Mittelwert u.Varianz bzw.Varianzkoeffizient) benötigt.

5.1 Einführung

Bei seriellen Systemen mit Blockierung (d.h. Blindbelegung einer BE in Stufe i bei voll belegter Stufe i+1) kann die maximale Durchsatzrate als wichtigste Verkehrsgröße des Systems angesehen werden. Diese ist unabhängig davon,bei welchem Ankunftsprozeß und welcher Speichergröße s_1 die Vollbelegung der 1.Stufe erreicht wurde.

Für das hier betrachtete allgemeine 2-stufige Single-Server System stellt

$$\rightarrow VF_1 \mid 1 \xrightarrow{s_2} VF_2 \mid 1$$

die in dieser Arbeit verwendete Symbolik dar,wobei VF_1 und VF_2 für die Verteilungsfunktionen der Bedienungszeiten stehen und s_2 die Größe des Zwischenspeichers angibt.

Für dieses System gibt es (vgl. Kap. 2) für den maximalen Durchsatz keine allgemeine Lösung,es existieren nur (exakte) Lösungen bzw.Lösungsverfahren für Grenzfälle der Speichergröße oder für Spezialfälle der **Kombinationen VF_1-VF_2 der Bedienungszeit-VF** (vgl.Tabelle 5.1)

Im Grenzfall des unendlich großen Zwischenspeichers handelt es sich um 2 Einzelstufen,wobei (vgl. 1.3.3)

$$\lambda_{max} = \min(\varepsilon_1, \varepsilon_2) \qquad (5.1)$$

während für $s_2 = 0$ eine exakte explizite Lösung von MAKINO angegeben wurde (vgl. 2.3.2.1),weshalb dieser Grenzfall im weiteren nur eine untergeordnete Rolle spielt,jedoch ist die nachfolgend entwickelte Näherung auch in diesem Falle anwendbar.

Veröffentlichung	s_2	VF_1-VF_2	Ergebnis für λ_{max}
——	∞	beliebig	trivial
MAKINO |82|	0	"	exakt
HUNT |73|	beliebig	M-M	exakt
AVI-ITZHAK |56|	"	D-D	exakt
HILLIER u. BOLING |72|	"	E_k-E_k	numerisch exakt
NEUTS |89|	"	G-M	prinzipielles, exaktes Verfahren
FINCH |34|	"	M-G	Analogie zu einstufigem System
Diese Arbeit	"	beliebig	approximativ

Tabelle 5.1 Ergebnisse für λ_{max} bei 2-stufigen Single-Server Systemen

Die angegebenen Spezialfälle für bestimmte VF_1-VF_2-Kombinationen wurden bis auf den Fall M-G in Abschnitt 2.3 erklärt, bei dem die Bestimmung des maximalen Durchsatzes auf die Berechnung einer einstufigen Anordnung zurückgeführt werden kann. Hierauf wird separat in 5.2 eingegangen.
Die in diesem Kapitel abgeleitete Näherung beruht auf einer durch Simulationsergebnisse ermittelten approximativ nach einem Exponentialgesetz beschreibbaren Abnahme des maximalen Durchsatzes mit der Summe der Quadrate der beiden Varianzkoeffizienten, wobei exakte Ergebnisse als Stützwerte dienen.

5.2 Prinzipielle Möglichkeit der Berechnung für M-G

Speziell bei negativ-exponentieller VF der Bedienungszeiten in Stufe 1 kann die Bestimmung der maximalen Durchsatzrate auf die Bestimmung von Größen in einem äquivalenten einstufigen System zurückgeführt werden. Diese Äquivalenz wurde von FINCH |34| bereits 1958 erwähnt, fand aber in der Literatur wenig Beachtung. Sie besagt, daß im 2-stufigen System mit M-G der maximale Durchsatz auch aus dem äquivalenten einstufigen System M|G|1-s* bestimmt werden kann, einer Einzelstufe mit Poisson-Ankünften der Rate

$$\lambda^* = \varepsilon_1 = \frac{1}{h_1} \quad \text{und} \quad s^* = s_2+1 \qquad (5.2)$$

Warteplätzen.

Dabei berücksichtigt der letzte Warteplatz Nr. s_2+1 den Effekt, daß bei Blockierung der 1.Stufe diese BE als zusätzlicher Warteplatz für Stufe 2 wirkt.
Zum Beweis der Äquivalenz sei auf Bild 5.1 verwiesen, wo der Ankunftsprozeß für Stufe 2 und der Poisson-Ankunftsprozeß der Einzelstufe einander gegenübergestellt werden.

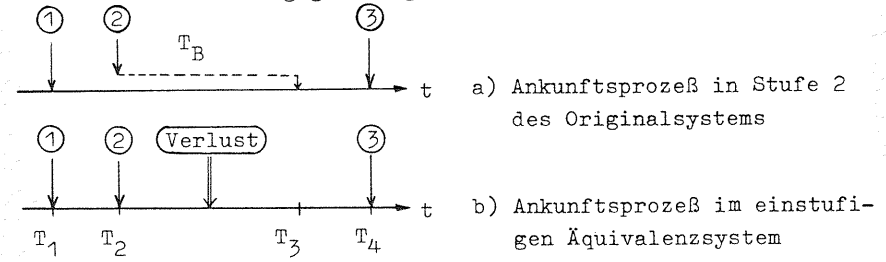


Bild 5.1a,b Gegenüberstellung der Ankunftsprozesse in beiden Systemen (Beispiel)

Solange die Zahl X_2 der Anforderungen in der 2. Stufe des Originalsystems kleiner ist als s_2+1, also noch mindestens ein freier Warteplatz vorhanden ist, gelangt im 2-stufigen System jede Anforderung ohne Blockierung in die 2.Stufe. Da die 1.Stufe immer voll belegt ist, sind dann beide Prozesse statistisch gesehen gleichwertig bzw. sogar identisch, wenn sie durch denselben Zufallsprozeß erzeugt werden. Belegt nun im Originalsystem eine Anforderung ① den letzten Warteplatz ($X_2=s_2+1$), dann kommt bei beiden Systemen die nächste Anforderung ② statistisch gesehen nach derselben Zeit (negativ-exponentiell verteilt mit Mittelwert h_1) an. Ist dann im 2-stufigen System der letzte Warteplatz noch belegt, wird diese Anforderung blockiert und der Poissonprozeß des Abfertigens in Stufe 1 für die Dauer der Blockierzeit T_B gestoppt, während im einstufigen Äquivalenzsystem die Anforderung ② den zusätzlichen letzten Warteplatz belegt.
Kommen nun während T_B im einstufigen System Anforderungen (Verlust) an, so gehen diese ohne Rückwirkung verloren. Eine weitere Anforderung ③ kommt im Originalsystem erst eine Zeit mit Mittelwert h_1 nach Ende der Blockierung, im einstufigen System statistisch gesehen zur gleichen Zeit T_4 an, da dort der Zeitpunkt T_3 (wie jeder Zeitpunkt eines Poissonprozesses) als Regenerationspunkt angesehen werden kann (der vergangene Prozeßverlauf unerheblich für die Zukunft ist).

Die im einstufigen Äquivalenzsystem durchgesetzte Rate λ_Y^* ist also identisch mit der maximalen Durchsatzrate des 2-stufigen Systems

$$\lambda_{max} = \lambda_Y^* = \lambda^*(1-B^*) \qquad (5.3)$$

mit B* als Verlustwahrscheinlichkeit der äquivalenten Einzelstufe, die man z.B. nach dem in 6.3 angegebenen Verfahren (über eine eingebettete Markoff-Kette) bestimmen könnte.
Es ist klar, daß diese Äquivalenz auch für $n_2>1$ gilt, jedoch auf VF_1=M beschränkt ist, da bei allgemeineren Prozessen nur bestimmte Zeitpunkte die Regenerationseigenschaft besitzen. Deshalb konnte diese prinzipielle Möglichkeit der Berechnung im vorliegenden allgemeineren Falle nicht wahrgenommen werden; jedoch wird in Kapitel 6 ein ähnliches (approximatives) Prinzip der Vereinzelung von Stufen entwickelt, das ebenfalls auf Systeme der Art M|G|1-s führt.
Es sei hier noch vorausschauend für Abschnitt 5.7 vermerkt, daß die Blockierwahrscheinlichkeit p_B für Anforderungen im Originalsystem identisch ist mit der Wahrscheinlichkeit, daß im einstufigen System eine erfolgreiche d.h. nicht abgewiesene Anforderung den letzten Warteplatz belegt:

$$P(\text{erfolgr. Anf. startet auf Platz } s_2+1) = \frac{p^*(s_2+1)}{1-B^*} = \frac{p^*(s_2+1)}{1-p^*(s_2+2)} \quad , (5.4)$$

während die Blockierzeit der Wartezeit dieser Anforderung auf diesem Warteplatz entspricht. Diese (Teil-)Wartezeit, die als Restbedienungszeit der gerade bedienten Anforderung aufgefaßt werden kann, ist aber im allgemeinen nicht einfach angebbar (vgl. 5.7).

5.3 Beweis einer Äquivalenz mit einem geschlossenen System

Das untersuchte allgemeine System, dessen symbolische Darstellung in 5.1 gezeigt wurde, kann auch wie in Bild 5.2a charakterisiert werden.

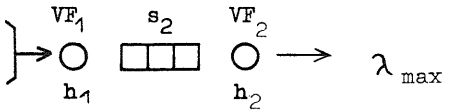


Bild 5.2a Untersuchtes System (mit Blockierung)

Dabei sind bei den BE zusätzlich mit VF_i der Funktionstyp und mit h_i der Mittelwert der beiden Bedienungszeit-VF angegeben (i=1,2), während der Zwischenspeicher mit der Zahl s_2 seiner Warteplätze versehen ist.
Hierzu kann man ein sog. geschlossenes System nach Bild 5.2b betrachten, in dem eine feste Anzahl N von Anforderungen zyklisch umläuft.

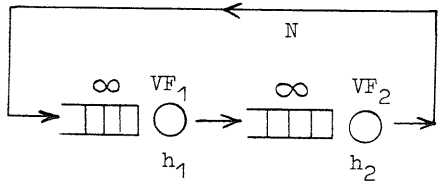


Bild 5.2b Geschlossenes System (ohne Blockierung)

Wegen der unendlich großen Wartespeicher ist dieses geschlossene System blockierungsfrei, die VF der Bedienungszeiten seien wie im offenen System nach Bild 5.2a.
Nun sei

$X_2(t)$ die Zahl der Anforderungen in Stufe 2 des offenen Systems zur Zeit t zuzüglich einer eventuell in Stufe 1 blockierten Anforderung

$X_2^*(t)$ die Zahl der Anforderungen in Stufe 2 des geschlossenen Systems zur Zeit t

SATZ 5.1 Die beiden Prozesse $X_2(t)$ und $X_2^*(t)$ sind für $N = s_2+2$ statistisch äquivalent bzw. identisch.

BEWEIS: Bei beiden Systemen sei der Zeitpunkt t=0 gekennzeichnet durch

$$X_2(0^-) = X_2^*(0^-) = s_2+2 \text{ bzw}$$

$$X_2(0^+) = X_2^*(0^+) = s_2+1$$

also
- im offenen System durch die Beendigung einer Blockierphase, hervorgerufen durch das Bedienungsende einer Anforderung in Stufe 2
- im geschlossenen System durch das Bedienungsende einer Anforderung in Stufe 2 bei noch leerer Stufe 1

Nun sei

$$A_2(t,x_2) \overset{def}{=} P(T_{A2} \leq t \mid X_2 = x_2) \quad \text{bzw.}$$

$$A_2^*(t,x_2) = P(T_{A2}^* \leq t \mid X_2^* = x_2)$$

die bedingte VF des Ankunftsabstandes T_{A2} bzw. T_{A2}^* in Stufe 2 des offenen bzw. geschlossenen Systems, beginnend mit dem Zeitpunkt des Entstehens des Zustandes $\{X_2\}$ bzw. $\{X_2^*\}$. Dabei zähle im offenen System eine blockierte Anforderung in Stufe 1 als "zu Blockierbeginn in Stufe 2 eigetroffen".
Nun gilt offensichtlich

$$A_2(t,x_2) = \begin{cases} H_1(\leq t) & \text{für } x_2 \leq s_2+1 \\ H_{2R}(\leq t) \ast H_1(\leq t) & \text{für } x_2 = s_2+2 \end{cases}$$

$$A_2(t,x_2) = \begin{cases} H_1(\leq t) & \text{für } x_2 \leq s_2+1 \\ H_{2R}^*(\leq t) \ast H_1(\leq t) & \text{für } x_2 = s_2+2 \end{cases}$$

Dabei sind $H_{2R}(\leq t)$ bzw $H_{2R}^*(\leq t)$ die VF der Restbedienungszeiten in Stufe 2, die im offenen System den Blockierzeiten T_B entsprechen, im geschlossenen System der (Teil-)Wartezeit auf Wartepatz Nr. s_2+1.
Diese beiden Zeiten ergeben sich -ausgehend z.B. von t=0 bei beiden Systemen- ausschließlich aus beiden VF VF_1 und VF_2, die jeweils für beide Systeme gleich sind. Deshalb gilt $H_{2R}(\leq t) = H_{2R}^*(\leq t)$ und somit sind Ankunftsprozeß und Bedienungsprozeß in beiden zweiten Stufen identisch. Daraus folgt SATZ 5.1.

Für den Fall rein negativ-exponentieller Bedienungszeiten wurde von GORDON und NEWELL |107| eine Äquivalenz zwischen obigem geschlossenen System und einem 1-stufigen System M|M|1-s_2+1 angegeben, die (vgl. Äquivalenz in 5.2) in Satz 5.1 enthalten ist. Dort wurde auch eine mögliche Erweiterung auf spezielle offene Systeme mit nicht-negativ-exponenetiellen VF angedeutet, was im Falle M-G bereits in 5.2 geschildert wurde (vgl. auch REISER und KOBAYASHI |54|).
Für VF_1 und $VF_2 \neq M$ würde jedoch Satz 5.1 auf eine äquivalente 1-stufige Anordnung führen, bei der aber der Ankunftsprozeß vom Bedienungsprozeß abhängen würde.

Aus Satz 5.1 folgt u.a., daß beide zweiten Stufen gleiche Belastung besitzen, also auch gleiche Zahl abgefertigter Anforderungen pro Zeiteinheit (durchgesetzte Rate, vgl. (1.40)).
Da aber beim blockierungsfreien geschlossenen System eine Reihenfolgevertauschung der beiden BE nur eine Umbenennung von Stufe 1 und 2 bedeutet, folgt hieraus

SATZ 5.2 Ein System von 2 BE in Serie mit beliebig verteilten Bedienungszeiten und begrenzter Zwischenspeichergröße $s_2 \geq 0$ besitzt unabhängig von der Reihenfolge der BE dieselbe maximale Durchsatzrate λ_{max}.

Somit läßt nun auch für 2-stufige Systeme mit G-M die maximale Durchsatzrate aus einem 1-stufigen System nach 5.2 berechnen.

Da die in der Literatur behandelten geschlossenen Systeme (vgl. REISER u. KOBAYASHI |54|) **mindestens** in einer Stufe eine negativ-exponentielle VF der Bedienungszeiten besitzen, konnten von von diesen Systemen keine Ergebnisse hier verwendet werden.

5.4 Untersuchte Systeme und Ergebnisse

Um den Anspruch der Anwendbarkeit des in diesem Kapitel entwickelten Berechnungsverfahrens belegen zu können, mußten, da es sich um ein Näherungsverfahren handelt, -das zudem noch bei sehr verschiedene VF und Systemparameternanwendbar sein soll- viele Kombinationen mit Hilfe der Simulation untersucht werden. Die hierbei gewonnenen Erkenntnisse und bestätigten Vermutungen wurden für das Näherungsverfahren selbst verwertet und werden ebenfalls in diesem Abschnitt geschildert.

Das in Bild 5.2a bereits dargestellte System besitzt 5 Parameter, die in den Simulationsreihen variiert wurden.

- Die untersuchten Zwischenspeichergrößen betrugen s_2=1,2,5 , was in Anbetracht des Single-Server Falles als ausreichend angesehen werden kann.

● Bei den VF der Bedienungszeiten wurden -wie bereits erwähnt- nur approximativ die ersten beiden Momente berücksichtigt.

• Als Zeiteinheit wurde -wie in dieser Arbeit üblich- die Summe aus den mittleren Bedienungszeiten gewählt:

$$h_1 + h_2 = 1 \text{ Z.E.} \qquad (5.5)$$

Deshalb war es ausreichend, die Werte $h_1 = 0.\overline{3}, 0.5, 0.\overline{6}$ zu betrachten, was ein Verhältnis $h_1/h_2 = 1/2, 1, 2$ bedeutet. Darüber hinaus wurden auch Werte von $h_1 = 0.1, 0.2$, bzw. $0.8, 0.9$ untersucht, die aber normalerweise in diesem Zusammenhang uninteressant sind, da dann der maximale Durchsatz praktisch ausschließlich von der langsameren BE bestimmt wird.

• Der in 1.2.1 eingeführte Varianzkoeffizient C einer VF ist die auf ihren Mittelwert bezogene Standardabweichung σ.

C_H^2	0	0.5	1.0	2.0
VF	D	E_2	M	H_2

Tabelle 5.2
Vorwiegend betrachtete VF

Tabelle 5.2 zeigt die hauptsächlich betrachteten Typen der Bedienungszeit-VF mit dem Quadrat ihrer Varianzkoeffizienten. Dabei sind für $C_H^2 < 1$ die VF vom Typ hypoexponentiell, während für $C_H^2 > 1$ eine hyperexponentielle VF (bezüglich der Varianz) vorliegt.

Dabei ist im Gegensatz zu einer Erlang-k-VF eine hyperexponentielle VF durch die Zahl ihrer (alternativen) Phasen noch nicht vollständig charakterisiert. Im vorliegenden Falle waren es 2 Zweige, die Mittelwerte und die Verzweigungswahrscheinlichkeiten wurden entsprechend des gewünschten Varianzkoeffizienten und einer weiteren Randbedingung (vgl. Anhang A1) bestimmt.
Für einige besondere Fälle wurden auch extreme Varianzen untersucht mit $C_H^2 \geq 8$, wie auch der Einfluß des 3. und höherer Momente (vgl. Anhang A2).

Die Ergebnisse aller Simulationstests wurden zunächst in 16 Tabellen angeordnet, von denen Tabelle 5.3 ein typisches Beispiel zeigt (Vertrauensintervalle nicht angegeben).

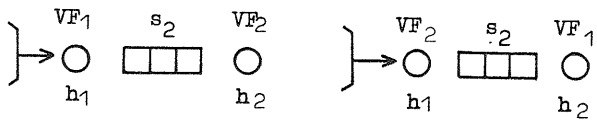
VF_1 \ VF_2	H_2	M	E_2	D	
H_2	1.47 0.25	1.53 0.31	1.59 0.34	1.60 0.40	λ_{max} p_B
M	1.51 0.21	1.60 0.25	1.67 0.27	1.73 0.30	λ_{max} p_B
E_2	1.58 0.17	1.67 0.21	1.71 0.22	1.84 0.21	λ_{max} p_B
D	1.61 0.13	1.74 0.16	1.84 0.13	2.00 0.00	λ_{max} p_B

Tabelle 5.3 Zusammenstellung der Simulationsergebnisse für ein typisches Beispiel ($h_1=h_2=0.5, s_2=2$)

Aus dem Vergleich der Werte für λ_{max} innerhalb und zwischen jeder der 16 Tabellen konnte folgender Satz abgeleitet werden:

SATZ 5.3 Bei vorgegebenen mittleren Bedienungszeiten in Stufe 1 und 2 und gegebener Zwischenspeichergröße s_2 ist der maximale Durchsatz mit guter Näherung unabhängig von der Vertauschung des Typs der VF in beiden Stufen.

Man erhält also für die beiden Systeme [1] und [2]



Ausgangssystem [1] System [2]

approximativ denselben maximalen Durchsatz.

Damit folgt mit Satz 5.2 resultierend:

SATZ 5.4 Alle 4 verschiedenen Systeme, die aus einer gegebenen Menge von 2 mittleren Bedienungszeiten und 2 Typen von VF gebildet werden können, besitzen approximativ die gleiche maximale Durchsatzrate λ_{max}.

Dieser Satz bildet eine gute Ausgangsbasis für die erwünschte, möglichst einfache aber doch brauchbare Näherung.

5.5 Analytische Untersuchung des aufgestellten Satzes

Die Gültigkeit des Satzes 5.4 bzw.,genauer gesagt,der Fehler in der maximalen Durchsatzrate bei Anwendung dieses Satzes wird im folgenden für verschiedene Zwischenspeichergrößen analytisch untersucht.

● Für $s_2 \to \infty$ verschwindet der Effekt der Blockierung und die maximal verarbeitbare Rate wird von der langsamsten Einzelstufe bestimmt.In diesem Falle sind alle $4\,\lambda_{max}$-Werte gleich (nach (5.1)),Satz 5.4 also mehr als gerechtfertigt.

● Für $s_2 = 0$ gilt nach MAKINO (vgl. 2.3.2.1):

$$\lambda_{max} = \frac{1}{E(\max\{T_{H1}, T_{H2}\})} \qquad (5.6)$$

Dabei ist

$$E(\max\{T_{H1}, T_{H2}\}) = \int_{t=0}^{\infty} t \cdot dH_{max}(\leqslant t) \qquad (5.7)$$

wobei $H_{max}(\leqslant t)$ die VF der Zufallsvariablen $\max\{T_{H1}, T_{H2}\}$ darstellt,für die aus der Wahrscheinlichkeitstheorie bekannt ist:

$$H_{max}(\leqslant t) = H_1(\leqslant t) \cdot H_2(\leqslant t) \qquad (5.8)$$

•$H_1(\leqslant t)$ und $H_2(\leqslant t)$ sind bezüglich λ_{max} vertauschbar,was einer Vertauschung der Reihenfolge der BE entspricht (Satz 5.2).

•Zur Beurteilung der Güte des Satzes 5.3 kann man die Durchsatzraten zweier entsprechender Systeme nach (5.6)-(5.8) exakt bestimmen und miteinander vergleichen.

Ein solches System ist z.B. sinnvollerweise ein System [a] mit M in Stufe 1 (Mittelwert h_1) und D in Stufe 2 (Mittelwert h_2), für das Satz 5.3 auf ein System [b] mit D in Stufe 1 (Mittelwert h_1) und M in Stufe 2 (Mittelwert h_2) führt.

System [a] :
Die VF der Bedienungszeiten lauten

$$H_1(\leqslant t) = 1 - e^{-t/h_1}, \quad H_2(\leqslant t) = \begin{cases} 0 & t<h_2 \\ 1 & t \geqslant h_2 \end{cases}$$

Hieraus folgt mit (5.8)

$$H_{max}(\leqslant t) = \begin{cases} 0 & t<h_2 \\ 1-e^{-t/h_1} & t \geqslant h_2 \end{cases}$$

Die zugehörige Dichtefunktion beträgt

$$f_{Hmax}(t) = \frac{dH_{max}(\leqslant t)}{dt} = \begin{cases} 0 & t<h_2 \\ \frac{1}{h_1} \cdot e^{-t/h_1} + (1-e^{-h_2/h_1}) \cdot \delta(t-h_2) & t \geqslant h_2 \end{cases}$$

Mit (5.7) ergibt sich

$$E(\max\{T_{H1}, T_{H2}\}) = \int_{t=h_2^-}^{h_2^+} t \cdot (1-e^{-h_2/h_1}) \cdot \delta(t-h_2)\,dt + \int_{t=h_2^+}^{\infty} \frac{t}{h_1} e^{-t/h_1}\,dt$$

Mit der sog. Ausblendeigenschaft der Dirac-Funktion $\delta(\cdot)$ ergibt sich schließlich für den maximalen Durchsatz im System [a]

$$\frac{1}{\lambda_{max}[a]} = h_2 + h_1 \cdot e^{-h_2/h_1} \qquad (5.9)$$

System [b] :
Für dieses System erhält man den maximalen Durchsatz wegen der Symmetrie in beiden VF durch Vertauschen von h_1 und h_2 in (5.9):

$$\frac{1}{\lambda_{max}[b]} = h_1 + h_2 \cdot e^{-h_1/h_2} \qquad (5.10)$$

Beide Kurven sind in Bild 5.3 als Funktion von $h_1/(h_1+h_2)$ einander gegenübergestellt.

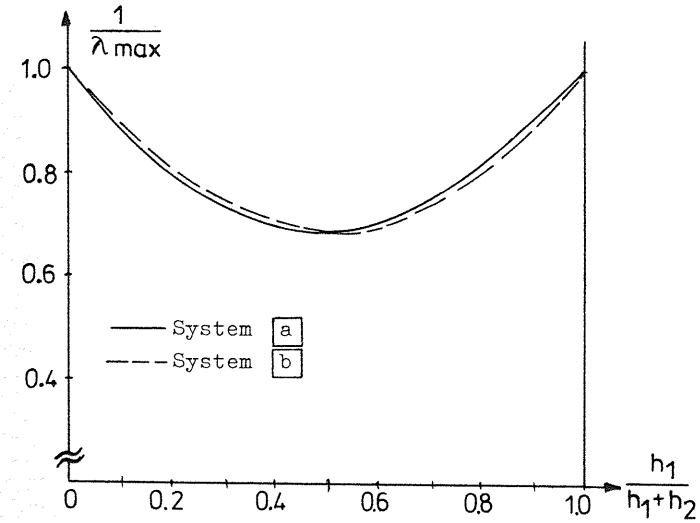


Bild 5.3 Vergleich der maximalen Durchsatzraten (s_2 =0)

Der Fehler bei beliebigem Verhältnis h_1/h_2 beträgt hier maximal nur 3%,kann also für die Praxis außer Betracht bleiben. Damit ist Satz 5.3 für diesen Fall gut bestätigt.

● Für endlich großen Zwischenspeicher $0<s_2<\infty$ ist bei $VF_1=VF_2$ aus der Literatur bekannt,daß λ_{max} für

M-M nach HUNT (vgl.2.3.1.1) und
D-D nach AVI-ITZHAK (vgl.2.3.2.2)

unabhängig von der Reihenfolge der Stufen ist (entspricht Satz 5.2).

Bei $VF_1\neq VF_2$ kann hierzu aus der Literatur nur das System mit G-M nach NEUTS herangezogen werden.Wendet man dessen Verfahren an,so kann man z.B. für den Fall D-M mit $s_2=1$ eine explizite Formel für λ_{max} herleiten und einen maximalen Fehler bei Vertauschen der mittleren Bedienungszeiten (Satz 5.1 und 5.3 kombiniert)berechnen,der bei beliebigem h_1/h_2 bereits <2% beträgt.

5.6 Gewählte Approximation

Aufgrund der Voruntersuchungen und der daraus abgeleiteten und in Satz 5.4 formulierten Ergebnisse ist es nun möglich, eine einfache Approximationsformel anzugeben.
Da bei festem h_1 und h_2 λ_{max} bei Vertauschung der beidenTypen der VF sich nur wenig ändert,kann der Einfluß der VF-Typen relativ gut durch einen Term beschrieben werden,der in den beiden Varianzkoeffizienten C_{H1} und C_{H2} symmetrisch ist. Dabei war es von der Anschauung her zunächst am sinnvollsten, hierzu die Varianz der gesamten Bedienungszeit einer Anforderung

$$\sigma_H^2=\sigma_{H1}^2+\sigma_{H2}^2 \qquad \text{mit} \quad \sigma_{Hi}=h_i\cdot C_{Hi} \quad i=1,2 \qquad (5.11)$$

zugrunde zu legen,was im vereinfachten Falle der Vertauschbarkeit von h_1 und h_2 auf den Term

$$C_{H1}^2+C_{H2}^2$$

führte,was durch die Simulationsergebnisse auch gut bestätigt wurde (vgl. Ergebniskurven in Bild 5.4-5.6).

Bild 5.4 zeigt den prinzipiellen Verlauf des maximalen Durchsatzes λ_{max} für gegebenes h_1 und h_2 als Funktion von $C_{H1}^2+C_{H2}^2$.

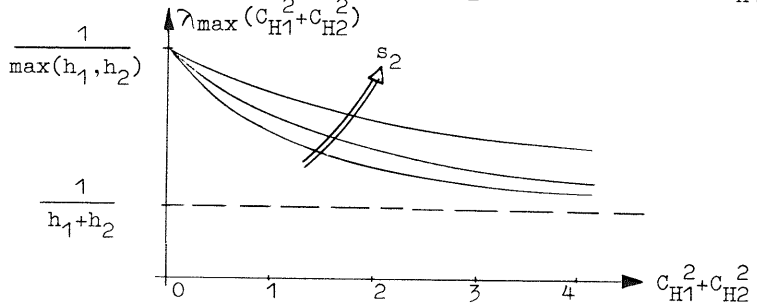


Bild 5.4 Prinzipieller Verlauf des maximalen Durchsatzes

Für $C_{H1}^2+C_{H2}^2=0$ sind die Bedienungszeiten in beiden Stufen konstant,also

$$\lambda_{max}(0)=\lambda_{maxD-D}=\frac{1}{\max(h_1,h_2)} \qquad (5.12)$$

Mit steigendem $C_{H1}^2+C_{H2}^2$ nimmt λ_{max} ab,kann aber nie unter den Grenzwert $1/(h_1+h_2)$ sinken,wo die Schwankungen der Bedienungszeiten so extrem sind,daß im Endeffekt nur 1 BE gleichzeitig bedient:

$$\lambda_{max}(\infty)=\frac{1}{h_1+h_2} \qquad (5.13)$$

Es liegt nun nahe,den Kurvenverlauf zwischen diesen beiden Extremwerten durch eine Exponentialfunktion anzunähern,wobei als Stützwert

$$\lambda_{max}(2)=\lambda_{maxM-M}=\frac{1-\left(\frac{h_1}{h_2}\right)^{s_2+2}}{1-\left(\frac{h_1}{h_2}\right)^{s_2+3}}\cdot\frac{1}{h_2} \qquad (5.14)$$

nach (2.10) sich anbot:

$$\lambda_{max}(C_{H1}^2+C_{H2}^2)=\lambda_{max}(\infty)+(\lambda_{max}(0)-\lambda_{max}(\infty))\cdot e^{-a\cdot f(C_{H1}^2+C_{H2}^2)} \qquad (5.15)$$

Dabei ist $f(\bullet)$ eine noch wählbare Funktion,die Konstante a wird so bestimmt,daß bei gewähltem $f(\bullet)$ (5.14) erfüllt wird. Durch Vergleich mit der Vielzahl der Simulationsergebnisse erwies sich die einfache Funktion

$$f(C_{H1}^2+C_{H2}^2)=\sqrt{C_{H1}^2+C_{H2}^2} \qquad (5.16)$$

als brauchbare Näherung.

Aus (5.12)-(5.16) folgt schließlich

$$\lambda_{max}(C_{H1}^2+C_{H2}^2)=\frac{1}{h_1+h_2}+\left(\frac{1}{\max(h_1,h_2)}-\frac{1}{h_1+h_2}\right)e^{-a\cdot\sqrt{C_{H1}^2+C_{H2}^2}}$$

wobei (5.17)

$$a=-\frac{1}{\sqrt{2}}\cdot\ln\left\{\frac{(h_1+h_2)\cdot\lambda_{maxM-M}-1}{\frac{h_1+h_2}{\max(h_1,h_2)}-1}\right\}$$

als explizite Näherungsformel für 2 BE in Serie mit den Varianzkoeffizienten C_{H1} und C_{H2}. Die Größe s_2 des Zwischenspeichers erscheint lediglich in λ_{maxM-M} nach (5.14); es ist direkt ersichtlich, daß auch der Fall $s_2=0$, für den bereits eine exakte Lösung existiert, mit einbezogen werden kann.

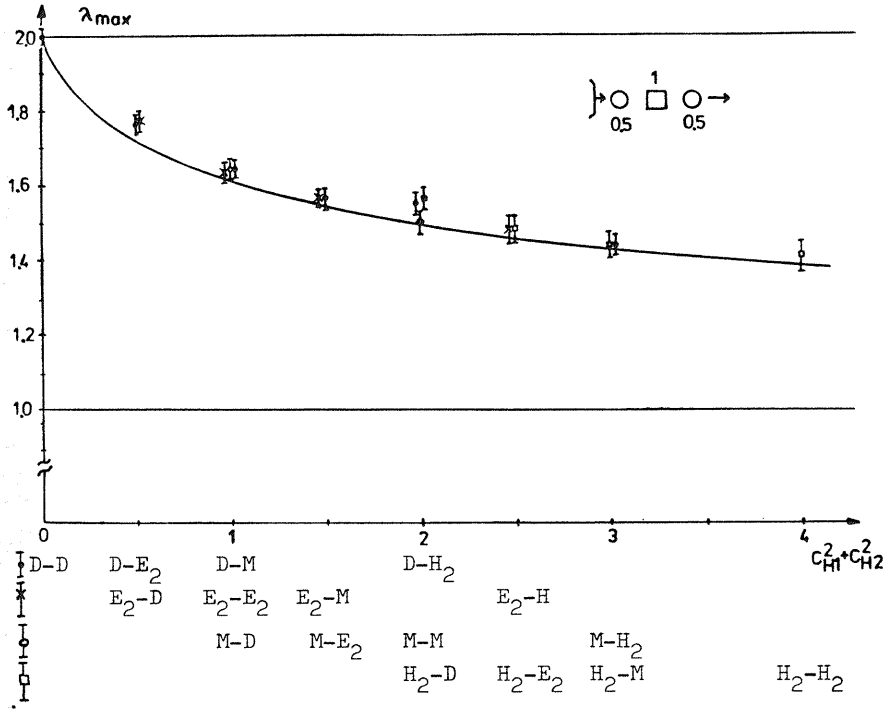


Bild 5.5 Maximale Durchsatzrate bei $h_1=h_2=0.5, s_2=1$

Die Bilder 5.5-5.7 zeigen typische Verläufe der Näherung nach (5.17) mit zugehörigen Ergebnissen der Simulation für 95%ige statistische Aussagesicherheit für die verschiedensten Bedienungszeit-VF-Kombinationen (vgl. Legende zu Bild 5.3).

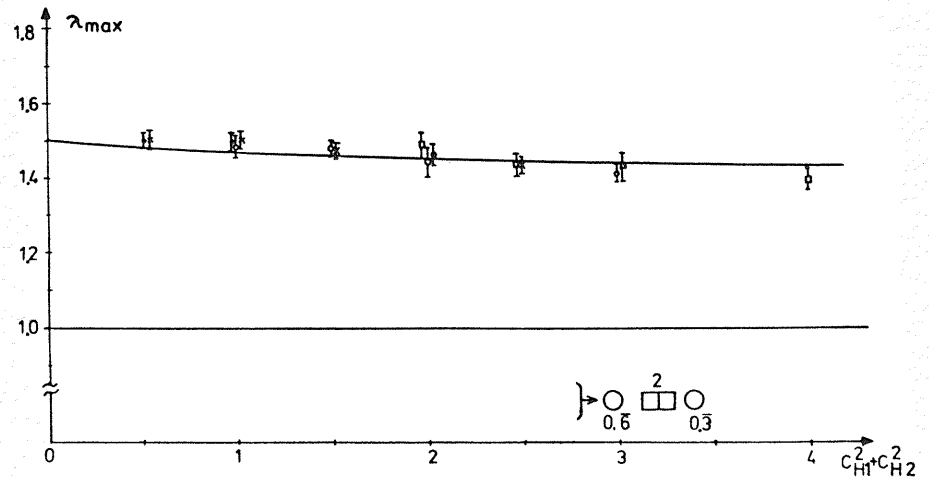


Bild 5.6 Maximale Durchsatzrate bei $h_1=0.\overline{6}, h_2=0.\overline{3}, s_2=2$

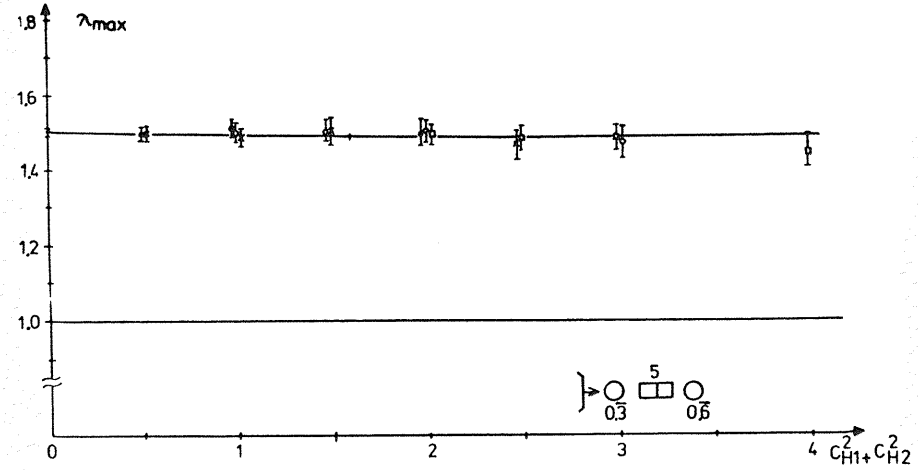


Bild 5.7 Maximale Durchsatzrate bei $h_1=0.\overline{3}, h_2=0.\overline{6}, s_2=5$

Auf den Einfluß des 3. und höherer Momente wird im Anhang A2 eingegangen. Systematische Diagramme mit s_2 als Parameter vgl. A3.

5.7 Mittlere Blockierzeit der Blockierten

In diesem Abschnitt wird für das betrachtete System zusätzlich eine Näherungsformel für die mittlere Blockierzeit der Blockierten (und damit auch der Blockierwahrscheinlichkeit) abgeleitet. Bei den VF wird eine Beschränkung auf hypoexponentielle Typen vorgenommen, die im Kapitel 6 ausschließlich betrachtet werden und wo diese Näherung für t_B auch für den Fall allgemeiner Durchsatzraten benötigt wird und zufriedenstellende Ergebnisse liefert.

Da bei maximalem Durchsatz die BE der 1.Stufe immer voll belegt ist (vgl. 1.3.2), gilt

$$\lambda_{max}\cdot(h_1+E(T_B)) = 1 \quad \text{bei} \quad \lambda_Y=\lambda_{max} \tag{5.18}$$

Kennt man also λ_{max}, ist damit auch die mittlere Blockierzeit aller Anforderungen

$$E(T_B) = p_B\cdot t_B \tag{5.19}$$

bekannt. Gelingt es nun, die mittlere Blockierzeit t_B der Blockierten anzugeben, so könnte damit auch ein Wert für die Blockierwahrscheinlichkeit p_B angegeben werden.

Günstig hierfür ist der Umstand, daß die durch Simulation erhaltenen Werte für t_B im Rahmen der hier interessierenden Zwischenspeichergrößen s_2=1,2,5 nur sehr schwach von s_2 abhängen (s_2=0 exakt angebbar).

Man hat nun zunächst die Möglichkeit, durch eine einfache, grobe Annahme einen Ausgangswert für t_B zu bestimmen, nämlich die, daß eine Blockierphase zu jedem Zeitpunkt gleichwahrscheinlich beginnt. Dann ist die mittlere Blockierzeit der Blockierten gleich der sog. mittleren Restbedienungszeit (mittlere (Teil-) Wartezeit einer Anforderung auf ihrem 1.eingenommenen Warteplatz) im reinen Wartesystem $M|VF_2|1$ (vgl.z.B. |6|):

$$t_B \approx \frac{1+c_{H2}^2}{2}\cdot h_2 \tag{5.20}$$

Für VF_2=M ergibt sich der exakte Wert $t_B=h_2$, für VF_2=D ist $t_B\approx h_2/2$ unabhängig von h_1 und VF_1.

Man kann nun leicht feststellen, daß dies nur eine grobe Näherung sein kann, insbesondere für D-D, wo für t_B exakt gilt

$$t_B = \begin{cases} 0 & h_2 \leqslant h_1 \\ h_2 - h_1 & h_2 \geqslant h_1 \end{cases} \tag{5.21}$$

Deshalb wurde ein anderer Weg zur approximativen Bestimmung von t_B eingeschlagen, der auf einer von den Varianzkoeffizienten abhängigen Interpolation zwischen teilweise exakten und noch herzuleitenden approximativen Werten beruht.

Tabelle 5.4 zeigt t_B für den Fall s_2=2 (nur geringe Abweichungen bei s_2=1,5) bei 3 verschiedenen Verhältnissen h_1/h_2.

t_B/h_2

VF_1 \ VF_2	M	E_2	D	h_1/h_2
M	1.00	0.80	0.63	0.5
	1.00	0.74	0.50	1
	1.00	0.75	0.39	2
E_2	1.00	0.80	0.57	0.5
	1.00	0.71	0.38	1
	1.00	0.69	0.21	2
D	1.00	0.73	0.50	0.5
	1.00	0.65	0	1
	1.00	—	0	2

Tabelle 5.4

(Normierte) Mittlere Blockierzeit der Blockierten (Werte für $VF_1\neq M$ aus Simulation, ohne Vertrauensintervalle; — nicht ermittelt, $p_B < 10^{-4}$).

Es wurden folgende Ansätze gemacht

$$t_{BM-VF2} = (1-c_{H2}^2)\cdot t_{BM-D} + c_{H2}^2\cdot h_2 \tag{5.22}$$

und

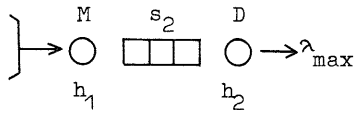
$$t_{BD-VF2} = (1-c_{H2}^2)\cdot t_{BD-D} + c_{H2}^2\cdot h_2 \tag{5.23}$$

Mit

$$t_{BVF1-VF2} = (1-c_{H1}^2)\cdot t_{BD-VF2} + c_{H1}^2\cdot t_{BM-VF2} \tag{5.24}$$

ist eine Möglichkeit der approximativen Bestimmung für hypoexponentielle VF gegeben.

Voraussetzung hierfür ist es, t_{BM-D} zu kennen. Hierzu sind in Bild 5.8 für das System



verschiedene Simulationsergebnisse für t_B als Funktion der mittleren Bedienungszeiten h_1 und h_2 bei verschiedenen Zwischenspeichergrößen aufgetragen.

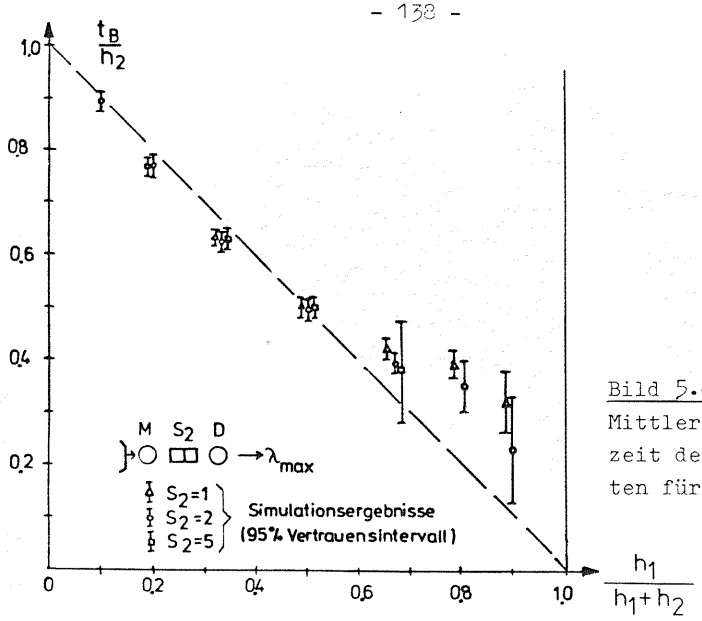


Bild 5.8
Mittlere Blockierzeit der Blockierten für M-D

Für $h_1 \to 0$ geht unabhängig von s_2 $t_B/h_2 \to 1$, da jede Anforderung in Stufe 1 sofort blockiert wird. Für $h_1 \to 1$, d.h. $h_2 \to 0$ bei $h_1+h_2=1$, wird t_B/h_2 immer kleiner, gleichzeitig aber sinkt die Blockierwahrscheinlichkeit stark, so daß t_B für diesen Fall relativ unkritisch ist. Deswegen wurde als einfache Näherung

$$\frac{t_{BM-D}}{h_2} = \frac{h_2}{h_1 + h_2} \qquad (5.25)$$

gewählt. Aus (5.21)-(5.25) ergibt sich nun für beliebige hypoexponentielle VF:

$$\frac{t_{BVF1-VF2}}{h_2} = (1-c_{H2}^2)\left\{(1-c_{H1}^2)\cdot \max\left[1-\frac{h_1}{h_2},0\right] + c_{H1}^2 \cdot \frac{h_2}{h_1+h_2}\right\} + c_{H2}^2 \qquad (5.26)$$

Für die untersuchten Systeme mit D, E_2, M ergeben sich für diese einfache Näherung folgende maximale Fehler

h_1/h_2	0.5	1	2
max. Fehler	<5%	<35%	<20%

Da dieses Ergebnis im nächsten Kapitel verwendet wird, sind hier keine weiteren Simulationsergebnisse für t_B gezeigt. Es sei diesbezüglich auf Kapitel 6 verwiesen, wo (5.26) auch für den Fall allgemeiner Durchsatzraten verwendet wird und ein Vergleich mit Simulationsergebnissen erfolgt.

6 ANWENDUNG EINES AUFSCHNEIDEPRINZIPS AUF 2-STUFIGE SINGLE - -SERVER SYSTEME MIT BLOCKIERUNG

In diesem Kapitel wird für das 2-stufige Single-Server System mit endlich großem Zwischenspeicher, für das in Kap.5 die maximale Durchsatzrate bestimmt wurde, ein Näherungsverfahren entwickelt zur Berechnung der Verkehrsgrößen bei allgemeiner Durchsatzrate.
Dabei wird der Fall des reinen Wartesystems betrachtet, mit einem Poisson-Ankunftsprozeß in Stufe 1 und beliebigen hypoexponentiell verteilten Bedienungszeiten in beiden Stufen.
Das Näherungsprinzip besteht darin, das 2-stufige System trotz Rückstaumöglichkeit in 2 quasi unabhängige Einzelstufen aufzuschneiden. Dabei ist das Charakteristikum des Ersatzmodells für Stufe 2 ein endlich großer Speicher, der deshalb entstehende Verlust wird in Beziehung gesetzt zur Blockierung im 2-stufigen Modell. Der Rückstau wird durch eine vergrößerte Belegungsdauer im Ersatzmodell für Stufe 1 berücksichtigt.
Wichtiges Ergebnis der Näherung ist die mittlere Gesamt-Verzögerungszeit von Anforderungen, aber auch Aussagen bezüglich der Verkehrsgrößen in den Einzelstufen.

6.1 Einführung

Serielle Wartesysteme mit endlich großen Zwischenspeichern sind durch den Effekt der Blockierung von Bedienungseinheiten (BE) gekennzeichnet (vgl.1.3). Dadurch verringert sich die mittlere Zahl der effektiv verfügbaren BE. Dies bedeutet, daß im Falle des reinen Wartesystems ($s_1 \to \infty$) die mittlere Wartezeit von Anforderungen in Stufe 1 bereits bei kleinerer Ankunftsrate eine Asymptote besitzt (Stationaritätsgrenze).
Die Lage der Asymptote, d.h. die maximale Durchsatzrate wurde für 2 BE in Serie mit endlich großem Zwischenspeicher in Kap.5 hergeleitet, in diesem Kapitel soll der Verlauf der mittleren Wartezeiten in Stufe 1 und anderer Größen bei allgemeinen Durchsatzraten bestimmt werden und zwar für den Fall des reinen Wartesystems mit einem Poisson-Ankunftsprozeß in Stufe 1.
Das betrachtete System kann also symbolisch dargestellt werden als

$$M \xrightarrow{\infty} VF_1|1 \xrightarrow{s_2} VF_2|1 \qquad (VF_1, VF_2 \text{ bel. hypoexp.}, \; 0<s_2<\infty)$$

Die Bedienungszeit-VF seien beliebig hypoexponentiell, wobei für die Näherung jedoch nur die ersten beiden Momente benötigt werden.
In den Abschnitten 2.3.1.2 und 2.3.2.2 wurde bereits eine ausführliche Übersicht über solche Systeme gegeben und insbesondere die in der Literatur behandelten Systeme mit den wesentlichsten Ergebnissen geschildert. Deshalb sind diese in Tabelle 6.1 nur kurz dargestellt.

Veröffentlichung	s_2	VF_1-VF_2	Ergebnisse
———	∞	beliebig	vgl. Tabelle 4.1
MORSE [86]	0	M-M	Gesamtzahl d. Anford. im System
SUZUKI [103] AVI-ITZHAK u. YADIN [56]	0	beliebig	viele Verkehrsgrößen
AVI-ITZHAK [57]	beliebig	D-D	Reduktion auf 1 Stufe
MAKINO [83]	"	M-M	Gesamtzahl d.Anford. im System (Approx.)
NEUTS [89],[90]	"	G-M	sehr komplex
Diese Arbeit	$0<s_2<\infty$	beliebig hypoexp.	viele Verkehrsgrößen (Approx.)

Tabelle 6.1 Behandelte 2-stufige Single-Server Systeme bei allgemeiner Durchsatzrate

Der Grenzfall des unendlich oder sehr großen Zwischenspeichers wurde eingehend in Kap. 4 beschrieben und wird hier nicht betrachtet, ebenso wie der Fall $s_2=0$, für den SUZUKI und AVI-ITZHAK eine ganze Reihe von Verkehrsgrößen bestimmten, die sich auf das Warten, die Blockierung und den Durchlauf von Anforderungen beziehen.
Neben diesen Grenzfällen für die Größe des Zwischenspeichers gibt es weitere Veröffentlichungen, die sich -bei beliebigem s_2- auf eine bestimmte Kombination der Bedienungszeit-VF beziehen. Hierzu zählt das System mit nur konstanten Bedienungszeiten, bei dem mit Hilfe der Reduktion auf eine äquivalente einstufige Anordnung exakte Aussagen über die gesamte Verzögerungs-bzw. Durchlaufszeit gemacht werden können, als auch Systeme mit nur negativ-exponentiellen VF, für die von MAKINO eine äquivalente approximative einstufige Anordnung angegeben wurde mit gleicher Varianz des Ausgangsprozesses.

Über die Ergebnisse von NEUTS für den Fall G-M schreibt BURKE [62] (vgl. 2.1), daß sie so kompliziert seien, daß ihre Brauchbarkeit in Frage gestellt werden muß. Deshalb wurde für die Untersuchungen in diesem Kapitel ein approximatives Verfahren gewählt, zumal das hier behandelte System gegenüber jenem eine beliebige hypoexponentielle VF in Stufe 2 besitzt.
In Tabelle 6.1 nicht aufgeführt sind Veröffentlichungen und Ergebnisse, die nur für den Fall der Grenzbelastung (maximaler Durchsatz) gelten. Bezüglich dieses Randfalles sei auf Kap. 5 verwiesen, wo diese Systeme (bzw. Betriebsweisen) behandelt werden.

6.2 Beschreibung des Näherungsprinzips

Durch den Effekt der Blockierung können die Belegungszustände in den beiden Stufen stark voneinander abhängig sein, deshalb müssen bei einer exakten Lösung beide Stufen gleichzeitig betrachtet werden, was im allgemeinen einen erheblichen Lösungsaufwand erfordert, sofern überhaupt dafür ein Lösungsverfahren bekannt ist.
Eine sinnvolle Forderung an ein Näherungsverfahren ist (sofern möglich) also das Zurückführen des 2-stufigen Systems auf 2 äquivalente Einzelstufen, wobei aber keineswegs der Effekt der Blockierung vernachlässigt werden kann.
In 5.2 wurde eine Methode geschildert, wie ein 2-stufiges System mit Blockierung auf ein äquivalentes 1-stufiges System zurückgeführt werden kann (vgl. dort). Diese Äquivalenz ist jedoch nur gültig für negativ-exponentielle VF in Stufe 1 und den Fall der Grenzbelastung $\lambda_Y = \lambda_{max}$ und konnte deshalb hier nicht verwendet werden. Dafür wurde ein (neues) Prinzip der Zurückführung auf Einzelstufen entwickelt, das auf dem Aufschneiden des 2-stufigen Systems in 2 äquivalente Einzelstufen beruht (vgl. Bild 6.1).

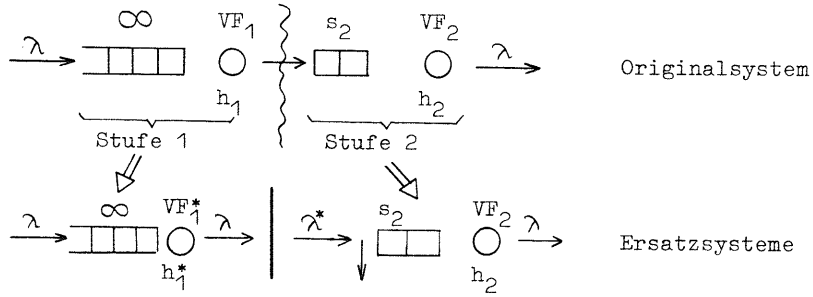


Bild 6.1 Abbildung des Originalsystems auf 2 Einzelstufen

● Als Ersatzsystem für Stufe 2 entsteht ein einstufiges Warte-Verlustsystem

$$VF^* | VF_2 | 1 - s_2$$

mit gleicher Bedienungszeit-VF, s_2 Warteplätzen und einem, hier noch nicht spezifizierten Ersatz-Ankunftsprozeß VF*, der die Zugangsverhältnisse zu Stufe 2 möglichst gut nachbilden soll. Die Ankunftsrate λ^* wird so gewählt, daß diese Einzelstufe die gleiche durchgesetzte Rate (Rate der erfolgreichen Anforderungen) und damit die gleiche Belastung besitzt wie die Originalstufe 2:

$$\lambda_Y^* = \lambda^*(1 - B^*) \stackrel{!}{=} \lambda_Y \qquad (6.1)$$

Dann kann angenommen werden, daß die gesuchten Verkehrsgrößen in Stufe 2 des Originalsystems (z.B. Wartewahrscheinlichkeiten, mittlere Wartezeiten) approximativ durch die entsprechenden auf die erfolgreichen Anforderungen bezogenen Größen des Ersatzsystems für Stufe 2 beschrieben werden können:

$$W_2 \approx W_{2Y}^* \qquad (6.2)$$

$$E(T_{W2}) \approx E(T_{W2})_Y^* \qquad (6.3)$$

Ebenfalls wird die Verlustwahrscheinlichkeit B* im Ersatzsystem in allererster Näherung gleich der Blockierwahrscheinlichkeit im 2-stufigen System gesetzt.
Es ist klar, daß insbesondere die letzte Annahme eine Näherung sein muß, da beide Effekte zwar ähnlich, aber doch verschieden sind. Während bei der Blockierung eine Anforderung (hier maximal eine) gewissermaßen vor Stufe 2 warten kann, geht diese im Ersatzmodell verloren. Die Güte dieser Näherungsannahme ist außerdem wesentlich von dem Ersatz-Ankunftsprozeß VF* abhängig, über den nun gesprochen werden soll.
Da in der Literatur über einstufige Wartesysteme bezüglich des obigen allgemeinen Systems mit endlich großem Wartespeicher (GI|G|1-s) keine hier verwendbaren Ergebnisse ermittelt werden konnten, wurde grundsätzlich als Ersatz-Ankunftsprozeß ein Poissonprozeß angenommen. Es wurde also zur Berechnung stets das System M|G|1-s herangezogen, das in 6.3 behandelt wird. Deshalb wurde als Ausgleich dafür die wichtigste Größe (p_B) über Adaptionsfaktoren ermittelt:

$$p_B \approx F_{VF_1 - VF_2} \cdot B^* \qquad (6.4)$$

Für diese Adaptionsfaktoren -deren Bestimmung in 6.5 erfolgt- konnten durchweg relativ einfache Ausdrücke gefunden werden, z.B. ist $F_{M-M} \approx 1$ (vgl. 6.5).

● Als Ersatzsystem für Stufe 1 entsteht durch Einbeziehung der Blockierzeiten in die Gesamtbelegungszeiten ein reines Wartesystem

$$M | VF_1^* | 1 ,$$

bei dem die Gesamtbelegungszeit einer Anforderung approximativ dargestellt wird durch eine Bedienungsphase entsprechend VF_1 und eine sich mit der Blockierwahrscheinlichkeit p_B daran anschließende Blockierphase mit Mittelwert t_B (mittl. Blockierzeit der Blockierten):

$$h_1^* = E(T_{H1}) + E(T_B) = h_1 + p_B \cdot t_B \qquad (6.5)$$

Als brauchbare Näherung für t_B konnte angenommen werden, daß dieser relativ unabhängig von der durchgesetzten Rate ist und also gleich dem Wert ist, der sich bei maximalem Durchsatz ergibt; dieser wurde in Abschnitt 5.5 Gleichung (5.26) hergeleitet (vgl. Simulationsergebnisse S.157)
Obige Annahme ist besonders deshalb möglich, da zur gleichen Zeit nur maximal eine Anforderung blockiert wird (n_1=1) und diese Größe sich nur auf blockierte Anforderungen bezieht, unabhängig davon, wie oft eine solche Blockierung auftritt.
Wegen Poissonankünften ist die Wartewahrscheinlichkeit im Ersatzsystem

$$W_1^* = \lambda \cdot h_1^* = A_1^* \qquad (6.6)$$

und es wird

$$W_1 \approx W_1^* \qquad (6.7)$$

gesetzt. Für die mittlere Wartezeit aller Anforderungen im Ersatzsystem für Stufe 1 ist nach POLLACZEK-KHINTCHINE (vgl.(4.1))

$$E(T_{W1})^* = \frac{\left(1 + \frac{\sigma_{H1}^{*2}}{h_1^{*2}}\right) \cdot A_1^*}{2(1 - A_1^*)} \cdot h_1^* \qquad (6.8)$$

und also nur von der gesamten Varianz σ_{H1}^{*2} von VF_1^* abhängig.
Die beiden möglichen Belegungsphasen einer Anforderung in diesem Ersatzsystem werden als unabhängig voneinander angenommen, deshalb gilt:

$$\sigma_{H1}^{*2} = \sigma_{H1}^{2} + \sigma_{B}^{2} \qquad (6.9)$$

Dabei ist σ_{H1}^2 die Varianz der eigentlichen Bedienungszeiten, σ_B^2 die Varianz der VF der Blockierzeiten bezogen auf alle Anforderungen.

Nimmt man nun an,daß die Blockierzeiten der Blockierten eine Varianz besitzen,die z.B. einer negativ-exponentiellen VF entspricht,so ist die Varianz der VF der Blockierzeiten bezogen auf alle Anforderungen gleich der Varianz einer speziellen hyperexponentiellen VF (die mit $1-p_B$ eine Blockierzeit =0 liefert) (vgl. Anhang A1):

$$\sigma_B^2 = (2p_B - p_B^2)\cdot t_B^2 \qquad (6.10)$$

Die gesuchte mittlere Wartezeit in Stufe 1 wird nun approximativ gleich der mittleren Wartezeit in diesem Ersatzsystem gesetzt:

$$E(T_{W1}) \approx E(T_{W1})^* \qquad (6.11)$$

Hierzu ist zu sagen,daß -selbst bei exaktem p_B und σ_B^2 - dies nur Näherungswerte liefert,da folgende 2 Unabhängigkeitsannahmen getroffen wurden,die aber nicht exakt erfüllt sind:

1) Die Bedienungszeit T_{H1} und die Blockierzeit T_B derselben Anforderung wie auch
2) die Gesamtbelegungszeiten aufeinenderfolgender Anforderungen seien voneinander unabhängig.

1) ist nicht erfüllt,da eine große zufällige Bedienungszeit T_{H1} die mittlere Blockierzeit dieser Anforderung herabsetzt.
2) ist sicher nicht exakt erfüllt,da im allgemeinen Korrelationen zwischen den Blockierzeiten aufeinanderfolgender Anforderungen existieren.

Beide Effekte sind jedoch in ihrer Tendenz auf die Wartezeiten in Stufe 1 einander entgegengesetzt,so daß sie sich mehr oder weniger kompensieren können.

6.3 Das Wartesystem M|G|1 mit endlichem Speicher

6.3.1 Bekannte Ergebnisse und gewählte Berechnungsmethode

Einstufige Wartesysteme mit endlich großem Speicher und Poissonankünften wurden bereits von ERLANG |3| betrachtet.Für das System M|M|n-s gab er mit Hilfe des Gleichungssystems für die stationären Zustandswahrscheinlichkeiten diese explizit an.Für den Fall n=1 lauten diese ($A=\lambda\cdot h$):

$$\left.\begin{aligned} &p(x) = p(0)\cdot A^x \qquad x=1..s+1 \\ \text{wobei}\quad &p(0) = \frac{1-A}{1-A^{s+2}} \end{aligned}\right\} \qquad (6.12)$$

Aus diesen ergeben sich weitere Verkehrsgrößen.Die Formeln werden hier angegeben,weil diese zur Berechnung der Verkehrsgrößen bei allgemeiner VF der Bedienungszeiten benötigt werden (Berechnung des Ersatzsystems für Stufe 2).

Wegen Poissonankünften ist die Verlustwahrscheinlichkeit

$$B = p(s+1) \qquad (6.13)$$

und die Belastung der BE

$$Y = A\cdot(1-B) \qquad (6.14)$$

Die Wartewahrscheinlichkeit (bezogen auf alle Anforderungen) beträgt

$$W = \sum_{i=0}^{s-1} p(i+1) = Y - B \qquad (6.15)$$

Im Ersatzsystem wird die Wartewahrscheinlichkeit bezogen auf alle erfolgreichen,d.h. nicht abgewiesenen Anforderungen benötigt,diese erhält man einfacherweise zu

$$W_Y = \frac{W}{1-B} \qquad (6.16)$$

Die Wartebelastung Ω ist die mittlere Länge der Warteschlange (ausschließlich der Anforderung in der BE)

$$\Omega = \sum_{i=0}^{s} i\cdot p(i+1) \qquad (6.17)$$

aus der direkt die mittlere Wartezeit bezogen auf alle erfolgreichen Anforderungen folgt:

$$E(T_W)_Y = \frac{\Omega}{\lambda\cdot(1-B)} = \frac{\Omega}{\lambda_Y} \qquad (6.18)$$

Schließlich gilt für die mittlere Wartezeit der Wartenden

$$t_W = \frac{E(T_W)_Y}{W_Y} \qquad (6.19)$$

Durch Einsetzen von (6.12) in (6.13)-(6.19) lassen sich im Falle negativ-exponentieller Bedienungszeiten einfache explizite Formeln angeben.

Für den Fall beliebiger VF der Bedienungszeiten (M|G|1-s) wurde von KEILSON |37| bewiesen,daß die stationären Zustandswahrscheinlichkeiten sich für x<s+1 von denen des zugehörigen reinen Wartesystems ($s \to \infty$) nur durch eine Konstante unterscheiden und deren Berechnung angegeben.Hiervon wird jedoch kein Gebrauch gemacht,da der Weg der Berechnung der stationären Zustandswahrscheinlichkeiten p(x) hier über die Zustandswahrscheinlichkeiten P_x an den Regenerationspunkten (Zeitpunkte des Abgangs von Anforderungen) einer eingebetteten Markoff-Kette führt (vgl. 6.3.2):

$$P(\text{erfolgr.Anf. lässt x Anf. im System zurück}) = P(X_D = x)_Y \stackrel{def}{=} P_x \qquad (6.20)$$

Diese Wahrscheinlichkeiten sollen separat in 6.3.2 durch Ansetzen der Markoff-Kette bestimmt werden (vgl. dort).

Es verbleibt also die Bestimmung der gewünschten stationären absoluten Zustandswahrscheinlichkeiten aus den -nun als bekannt angenommenen- Wahrscheinlichkeiten an den Regenerationspunkten. Hierzu kann man ein wichtiges Theorem über die Gleichheit der Zustandswahrscheinlichkeiten zu Ankunfts-und Abgangszeitpunkten eines Systems benutzen (vgl. |6|),aufgrund dessen hier für die Zahl X_A der von erfolgreichen Anforderungen angetroffenen Anforderungen im System gilt:

$$P(X_A = x)_Y = P(X_D = x)_Y \qquad x=0..s \qquad (6.21)$$

Dies bedeutet,daß erfolgreiche Anforderungen mit gleicher Wahrscheinlichkeit x Anforderungen im System antreffen (Antreffwahrscheinlichkeit),wie sie bei Verlassen des Systems x Anforderungen zurücklassen.

Da aber der Ankunftsprozeß ein Poissonprozeß ist,müssen diese Antreffwahrscheinlichkeiten der Erfolgreichen bekanntermaßen identisch sein mit den bedingten stationären Zustandswahrscheinlichkeiten,bezogen auf den Zeitraum,in dem das System

nicht ganz voll ist (Anforderungen erfolgreich sein können):

$$P(X_A = x)_Y = P(X=x \mid X<s+1) \qquad (6.22)$$

Für diese Antreffwahrscheinlichkeiten gilt

$$P(X=x \mid X<s+1) = \frac{P(X=x, X<s+1)}{P(X<s+1)} = \frac{p(x)}{1-p(s+1)} \qquad x=0..s \qquad (6.23)$$

Damit folgt aus (6.20)-(6.23):

$$p(x) = (1-B) \cdot P_x \qquad x=0..s \qquad (6.24)$$

Als einzige stationäre Zustandswahrscheinlichkeit fehlt noch p(s+1)=B,die man aus (6.14) mit einer weiteren Beziehung für die Belastung

$$Y = 1 - p(0) \qquad (6.25)$$

gewinnt zu

$$p(s+1) = 1 - \frac{1}{A+P_0} \qquad (6.26)$$

Damit ist es mit Hilfe von (6.24) und (6.26) sehr einfach,bei bekannten Zustandswahrscheinlichkeiten an den Regenerationspunkten die stationären Zustandswahrscheinlichkeiten zu beliebigen Zeitpunkten zu bestimmen.

Um den gleichen Durchsatz im Ersatzsystem für Stufe 2 wie im 2-stufigen System zu erhalten,kann eine programmtechnisch einfache Iteration des Angebots A erfolgen,die nur wenige Iterationsschritte benötigt (<15 für Restfehler $< 10^{-6}$).

6.3.2 Bestimmung der Zustandswahrscheinlichkeiten an den Regenerationspunkten

In 6.3.1 wurde die Berechnung der absoluten Zustandswahrscheinlichkeiten p(x) im Wartesystem M|G|1-s auf die Bestimmung der Zustandswahrscheinlichkeiten P_x des Systems zurückgeführt,die für Zeitpunkte gelten,in denen eine Anforderung gerade das System verlassen hat.

Da die Zahl der Anforderungen,die von einer Test-Anforderung im System zurückgelassen werden nur davon abhängt,wieviele Anforderungen ihre Vorgängerin zurückgelassen hat und der Zahl der während der Bedienungszeit der Test-Anforderung angekommenen Anforderungen,bilden diese Zustände die Regenerationspunkte einer eingebetteten Markoff-Kette (vgl. 1.2.3).Bezeichnet man mit q_i die Wahrscheinlichkeit,daß während einer zufälligen (entsprechend G verteilten) Bedienungszeit genau i Anforderungen in

einem Poissonprozeß mit der Rate λ ankommen, so lautet das Gleichungssystem für die Regenerationspunktwahrscheinlichkeiten P_x $(0 \leqslant x \leqslant s, s>0)$ (leere Summen ≡ 0):

$$P_x = (P_o + P_1)\cdot q_x + \sum_{i=2}^{x+1} P_i \cdot q_{x+1-i} \qquad 0 \leqslant x < s \qquad (6.27a)$$

$$P_s = (P_o + P_1)\cdot Q_s + \sum_{i=2}^{s} P_i \cdot Q_{s+1-i} \qquad (6.27b)$$

mit

$$Q_\nu = \sum_{i=\nu}^{\infty} q_i = 1 - \sum_{i=0}^{\nu-1} q_i \qquad \nu > 0 \qquad (6.28)$$

Eine Test-Anforderung läßt z.B. genau 1 Anforderung im System zurück (x=1), wenn während ihrer Bedienungszeit entweder genau 1 Anforderung ankam und die Vorgängerin keinen oder höchstens die Test-Anforderung im System zurückließ, oder keine Anforderung ankam und die Vorgängerin genau 2 Anforderungen zurückließ. Gleichung (6.27b) stellt einen Randfall dar und ist in den anderen s Gleichungen enthalten. An ihre Stelle tritt die Normierungsbedingung

$$\sum_{x=0}^{s} P_x = 1 \qquad (6.29)$$

Dieses Gleichungssystem wurde auch bei FINCH [34] gefunden, der als Lösung die erzeugende Funktion für die P_x angibt, von der durch Reihenentwicklung die einzelnen P_x berechnet werden müssten. Dieses in diesem Falle sehr aufwendige Verfahren wurde nicht gewählt, sondern ein für die Programmierung sehr geeignetes rekursives Schema.
Durch Umschreiben kann (6.27a) in folgende rekursive Form gebracht werden:

$$\left.\begin{aligned} P_1 &= -\frac{1}{q_o}\left\{ P_o \cdot q_o - P_o \right\} \\ P_x &= -\frac{1}{q_o}\left\{ (P_o+P_1)\cdot q_{x-1} - P_{x-1} + \sum_{i=2}^{x-1} P_i \cdot q_{x-i} \right\} \qquad 1 < x \leqslant s \end{aligned}\right\} \qquad (6.30)$$

Bei der Programmierung wird P_o zunächst ein beliebiger Wert, z.B. =1, zugewiesen und jede Gleichung nur einmal verwendet. Dann werden alle P_x entsprechend (6.29) normiert.
Die Berechnung der hier benötigten q_i wird im Anhang A1 geschildert, wo diese für die verschiedenen VF angegeben werden.

6.4 Untersuchte Parameterbereiche und Simulationsreihen

6.4.1 Verteilungsfunktionstypen

Wie in Kap.5 werden auch hier bei den VF nur die ersten beiden Momente berücksichtigt, was durch Angabe des Mittelwerts und des Varianzkoeffizienten erfolgt. Aufgrund der sehr großen Zahl von Parametern wurden die Betrachtungen auf beliebige hypoexponentielle VF beschränkt, so daß gilt:

$$c_{Hi}^2 = \sigma_{Hi}^2 / h_i^2 \leqslant 1 \qquad (i=1,2)$$

Dies ist ein Varianzspektrum, das z.B. von einer E_k-VF mit $c_H^2 = 1/k$ (k=1,2..) nur unvollständig wiedergegeben würde (vgl. Anhang A1). Es wurden zunächst die Typen M, E_2, D (c_H^2=1, 0.5, 0) untersucht, zur Entwicklung des Verfahrens. Dabei wurden die für die Berechnung des Ersatzsystems für Stufe 2 notwendigen Antreffwahrscheinlichkeiten q_i entsprechend einer im Anhang A1 beschriebenen "verschobenen negativ-exponentiellen VF (DM)" berechnet. Damit können alle Varianzkoeffizienten $0 \leqslant c_H \leqslant 1$ realisiert werden, die Fälle D und M sind als Randfälle enthalten.
Dabei ist zur Kombination D-D zu sagen, daß diese Systeme bezüglich der mittleren Gesamtzeit für Durchlauf und Verzögerung sehr einfach exakt zu berechnen sind (vgl. 2.3.2.2). Die Ergebnisse des Näherungsverfahrens für kleinere Varianzen sind jedoch umso besser, je besser die Ergebnisse des Näherungsverfahrens in diesem Randfalle sind. Deshalb war die Betrachtung von Systemen mit D-D für die Entwicklung des Näherungsverfahrens notwendig.
Darüber hinaus wurden weitere VF-Typen betrachtet, wie verschiedene Mehrpunkt-VF, als auch VF mit nicht durch E_k darstellbaren Varianzkoeffizienten. Diese wurden im Zusammenhang mit dem Einfluß des 3. und höherer Momente untersucht und geschlossen im Anhang A2 dargestellt.

6.4.2 Verhältnisse der mittleren Bedienungszeiten

Wie bei allen Untersuchungen wurde hier $h_1+h_2=1$ Zeiteinheit gewählt. Es wurde der Bereich

$$1/2 \leqslant h_1/h_2 \leqslant 2$$

untersucht. Bei außerhalb diesem Bereich liegenden Werten kann die gesamte Durchlaufszeit relativ gut mit Hilfe sog. "Flaschen-

hals-Modelle" bestimmt werden:

- Für $h_1/h_2>2$ liegt der Engpaß eindeutig in Stufe 1,Stufe 2 kann wegen"kleiner Blockierung" praktisch vergessen werden.
- Für $h_1/h_2<1/2$ wirkt die BE der Stufe 1 hauptsächlich als zusätzlicher Warteplatz für Stufe 2 (sie ist meist blockiert oder frei). Damit kann das ganze System mit guter Genauigkeit als Einzelstufe $M|VF_2|1$ berechnet werden.

6.4.3 Größe des Zwischenspeichers

Da hier nur Single-Server Stufen betrachtet wurden,war es ausreichend,die Zwischenspeichergrößen $s_2=1,2,5$ zu betrachten. Der Fall $s_2=0$ kann exakt berechnet werden.Für $s_2>5$ ist die Blockierung im Rahmen der betrachteten Varianzen der Bedienungszeiten sehr klein,Stufe 1 kann dann als Einzelstufe berechnet werden und Stufe 2 nach dem Näherungsverfahren in Kapitel 4.

6.4.4 Ankunftsraten

Aufgrund der gewählten Normierung und des Parameterbereichs für h_1 und h_2 würde sich bei unendlich großem Zwischenspeicher $1.5\leqslant\lambda_{max}\leqslant2.0$ ergeben.Für die untersuchten Systeme mit Blockierung waren also meist die Werte $\lambda=0.\overline{3},0.\overline{6},1.0,1.\overline{3}$ ausreichend.Zusätzlich konnten die Simulationsergebnisse für $\lambda_Y=\lambda_{max}$ aus Kap.5 herangezogen werden.

6.4.5 Simulationsreihen

Wegen der großen Zahl von Parametern (Systemen) und der Ergebnisgrößen pro System mußte die Überprüfung bzw. Entwicklung des Verfahrens anhand gezielter Untersuchungen erfolgen. Dazu wurden jeweils für die beiden Fälle $VF_1=D$ und $VF_1=M$ folgende Simulationsreihen durchgeführt:

I)	Variation von	VF_2	beim "Arbeitspunkt"		$s_2=2,h_1=0.5$
II)	"	" s_2	"	"	$VF_2=M,h_1=0.5$
III)	"	" h_1	"	"	$s_2=2,VF_2=E_2$

Weitere Simulationsreihen werden im Anhang A2 geschildert,wo auch andere VF-Typen und der Einfluß des 3. und höherer Momente untersucht werden.

6.5 Bestimmung der Adaptionsfaktoren

Bei dem in 6.2 erklärten Näherungsprinzip war -wegen einheitlicher Annahme eines Poisson-Ersatzankunftsprozesses in Stufe 2- die Einführung der Adaptionsfaktoren $F_{VF1-VF2}$ notwendig. Diese Faktoren werden hier für die verschiedenen Systeme bestimmt und können -was als Glücksfall angesehen werden mag- für die verschiedensten Fälle durch relativ einfache Ausdrücke approximiert werden.Es werden zunächst alle 4 Bedienungszeit-VF-Kombinationen betrachtet,die nur M und/oder D enthalten.Für die Adaptionsfaktoren beliebiger (hypoexponentieller) VF-Kombinationen erfolgt eine Interpolation entsprechend der Varianzkoeffizienten.

Im Falle negativ-exponentieller VF in beiden Stufen (M-M) fanden Voruntersuchungen statt für den -hier nicht betrachteten- Fall $s_2=0$,der exakt nach AVI-ITZHAK berechenbar ist.Berechnet man hiernach die Blockierwahrscheinlichkeiten bei fester Belastung der 2.Stufe als Funktion des Verhältnisses h_1/h_2,so ist im interessierenden Bereich $0.5\leqslant h_1/h_2\leqslant2$ -bei zusätzlicher Berücksichtigung der Stationaritätsgrenzen- p_B nur geringfügig ($<\pm7\%$) von h_1/h_2 abhängig.Dies heißt,daß bei M-M selbst für den härtesten Fall $s_2=0$ die Blockierwahrscheinlichkeit relativ unabhängig von der ersten Stufe ist,und daß damit für $s_2>0$ ein von h_1/h_2 unabhängiger Adaptionsfaktor F_{M-M} denkbar ist.

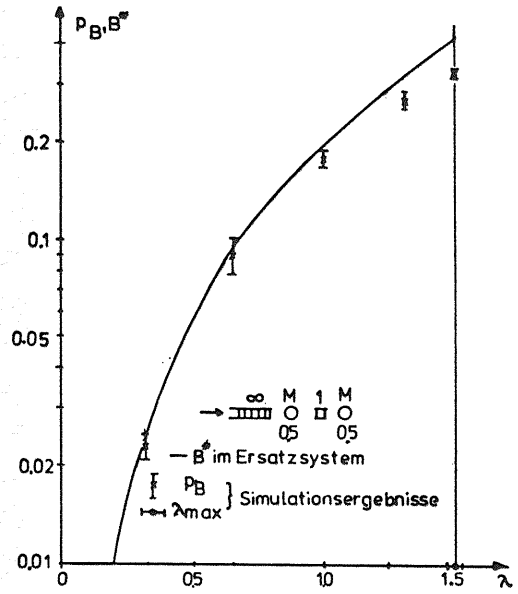


Bild 6.2
Vergleich Blockierung - Verlust bei M-M

Bild 6.2 zeigt zunächst einen für M-M typischen gegenseitigen Verlauf der Blockierwahrscheinlichkeit p_B (Simulationsergebnisse) und der Verlustwahrscheinlichkeit B* im Ersatzsystem für Stufe 2 über der Ankunftsrate λ des 2-stufigen Systems.
Durch systematische Untersuchung dieser p_B-B*-Relationen durch die Simulation und durch geeignete Wahl des Auftragens wurden nun die Faktoren $F_{VF1-VF2}$ bestimmt. Dabei war es von großem Vorteil, diese Faktoren über einer normierten Ankunftsrate λ/λ_{max} aufzutragen, wobei λ_{max} die maximal vom jeweiligen System verarbeitbare Rate darstellt, für die der sehr einfach berechenbare Wert nach dem Näherungsverfahren in Kapitel 5 genommen werden konnte.

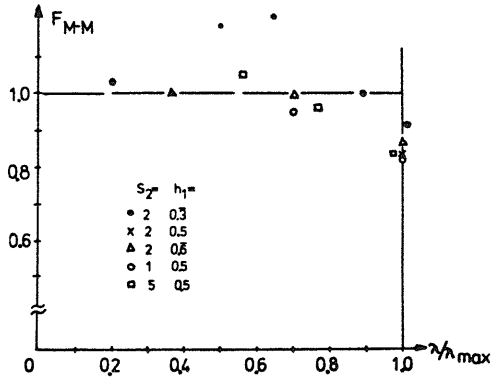


Bild 6.3
Adaptionsfaktoren F_{M-M}

Bild 6.3 zeigt diesen Verlauf für den Fall M-M bei verschiedenem h_1 und verschiedenen Zwischenspeichergrößen s_2. Dieser wurde durch Vergleich mit Simulationsergebnissen gewonnen, wobei in dieser Darstellung keine Vertrauensintervalle berücksichtigt wurden. Wie man aus den Ergebnissen für h_1=0.5 deutlich erkennen kann, besteht bei dieser Art des Auftragens keine spürbare Abhängigkeit von der Größe s_2 des Zwischenspeichers. Auch ist hier h_1 bzw. h_1/h_2 ohne größeren Einfluß, so daß

$$F_{M-M} \approx 1 \qquad (6.31)$$

als akzeptable Näherung im Bereich $0<\lambda/\lambda_{max}<0.9$ gelten kann.

Für die Fälle M-D und D-M traten, wie zumindest für D-M zu erwarten war, größere Unterschiede zwischen p_B und B* auf, vgl. Bilder 6.4 und 6.5.

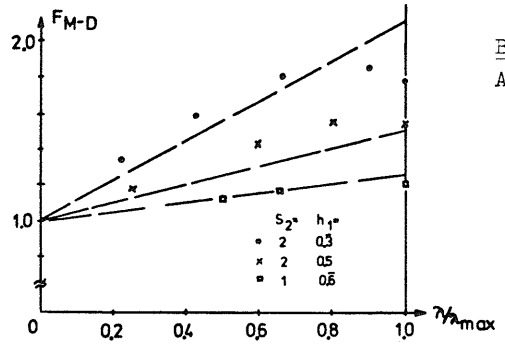


Bild 6.4
Adaptionsfaktor F_{M-D}

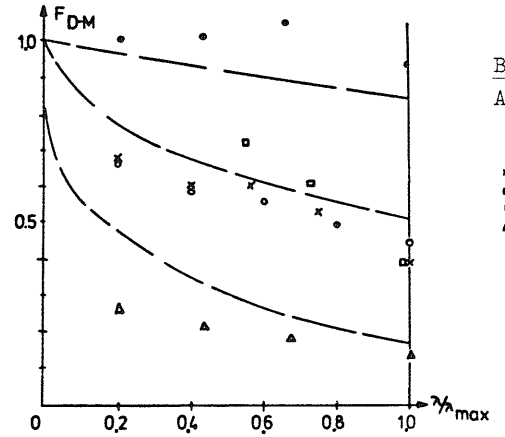


Bild 6.5
Adaptionsfaktor F_{D-M}

Auch hier ergaben sich keine spürbaren Abhängigkeiten von s_2, so daß die einfachen empirischen Approximationen

$$F_{M-D} = 1+\frac{\lambda}{\lambda_{max}}\cdot\frac{h_2}{2h_1} \qquad (6.32) \qquad F_{D-M} = 1-\left(\frac{\lambda}{\lambda_{max}}\right)^{\frac{h_2}{2h_1}}\cdot\left(\frac{1}{2}-\frac{h_2-h_1}{h_1+h_2}\right) \qquad (6.33)$$

zufriedenstellende Ergebnisse lieferten. Dabei ist bei kleinem p_B/B^* zu beachten, daß bei Berücksichtigung der Vertrauensintervalle die eingezeichneten Mittelwerte mit einer Ungenauigkeit bis maximal ± 30% behaftet sind.
Die größten Unterschiede ergaben sich natürlich für den Grenzfall D-D. Zunächst einmal ist trivialerweise für $h_1 \geq h_2$ dieser Faktor =0, da dort weder eine Anforderung im Zwischenspeicher wartet, noch

Blockierung stattfindet.
Für $h_1<h_2$ ergab sich bei $h_1=0.\overline{3}$ (vgl. Bild 6.6) ein stärkerer Einfluß der Zahl der Warteplätze im Zwischenspeicher, der erwartungsgemäß für $\lambda/\lambda_{max}\to 1$ wieder verschwindet, weil dort unabhängig von s_2 $p_B\to 1$ geht und ebenfalls wegen voller 2.Stufe $B^*\to 1$ gehen muß.

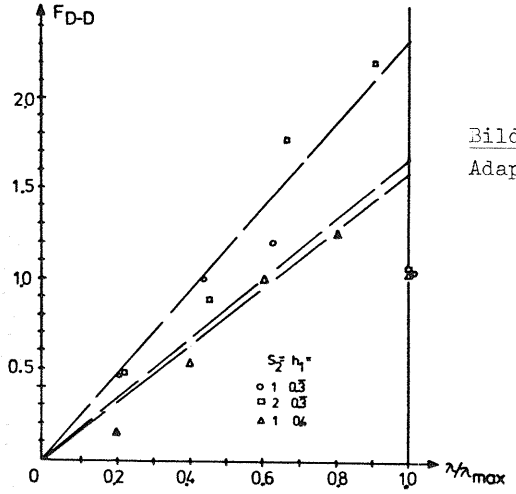


Bild 6.6
Adaptionsfaktor F_{D-D}

Für den Bereich $\lambda/\lambda_{max}<0.9$ wurde

$$F_{D-D} = \begin{cases} 0 & h_1 \geq h_2 \\ \frac{\lambda}{\lambda_{max}} \cdot (1+ s_2 \cdot \frac{h_2}{h_1+h_2}) & h_1<h_2 \end{cases} \qquad (6.34)$$

gewählt, wobei sich für $h_1=h_2$ eine Unstetigkeit ergibt, die allerdings auch im 2-stufigen System bei p_B vorhanden ist.
Der Adaptionsfaktor für beliebige hypoexponentielle VF-Kombinationen ergibt sich nun durch die gleiche Gewichtung wie in Kap.4:

$$F_{VF1-VF2} = C_{H1}^2 \cdot C_{H2}^2 \cdot F_{M-M} + (1-C_{H1}^2) \cdot C_{H2}^2 \cdot F_{D-M} + C_{H1}^2 \cdot (1-C_{H2}^2) \cdot F_{M-D} + (1-C_{H1}^2)(1-C_{H2}^2) \cdot F_{D-D} \qquad (6.35)$$

Dabei werden die Teilgrößen nach (6.31-6.34) berechnet, wobei als λ/λ_{max} die normierte Ankunftsrate des allgemeinen Systems herangezogen wurde, mit λ_{max} approximativ nach (5.17).

6.6 Typische Ergebnisse und Güte des Verfahrens

Ziel der Näherung war es, ein numerisch schnelles Verfahren zu entwickeln, das für die verschiedensten Parameter innerhalb der gesteckten Parameterbereiche (vgl.6.4) zufriedenstellende Ergebnisse liefert. Dabei war vor allem die gesamte Verzögerungszeit $E(T_{WB})$ von Interesse, die sich als Summe der mittleren Wartezeiten in beiden Stufen und der mittleren Blockierzeit in Stufe 1 bezogen auf alle Anforderungen ergibt:

$$E(T_{WB}) = E(T_{W1}) + E(T_B) + E(T_{W2})$$

Da in vielen betrachteten Fällen $E(T_{W2}) <$ bzw $\ll E(T_{W1})$ ist, war $E(T_{W2})$ selbst nicht von besonderem Interesse. Dagegen war es wegen der Blockierung und ihres starken Einflusses auf $E(T_{W1})$ notwendig, $E(T_B)=p_B \cdot t_B$ genauer zu kennen, wozu die in Kap.5 hergeleitete approximative Formel (5.26) für t_B mit herangezogen wurde.
Wegen des Aufschneidens in 2 äquivalente Einzelstufen ist es möglich, auch die Wartewahrscheinlichkeiten W_1 und W_2 und damit die mittleren Wartezeiten aller Wartenden t_{W1} und t_{W2} approximativ anzugeben, für Stufe 1 meist mit guter Genauigkeit, für Stufe 2 als gröbere Näherung. Diese spezielleren Größen sollen hier jedoch nicht weiter verfolgt werden.
Aufgrund der Vielzahl der Parameter sowie der Ergebnisgrößen können hier nur einige exemplarische Fehler innerhalb des untersuchten Spektrums angegeben werden, das durch die in 6.4 dargestellten Simulationsreihen und weitere Untersuchungen in etwa überdeckt wurde.
Die Bilder 6.7-6.9 zeigen am Beispiel des Systems mit M-D, $s_2=2$ und $h_1=h_2=0.5$ den Verlauf der Näherungswerte im Vergleich mit Simulationsergebnissen (Vertrauensintervalle mit 95% statistischer Aussagesicherheit).
Größere Abweichungen ergeben sich dabei nur für $E(T_{W2})$, was vermutlich darauf zurückzuführen ist, daß die BE der 1.Stufe zeitweilig als zusätzlicher Warteplatz für Stufe 2 wirkt und so Anforderungen zum frühest möglichen Zeitpunkt in Stufe 2 gelangen (Nachschubeffekt), was im Ersatzankunftsprozeß für Stufe 2 nicht berücksichtigt wurde.

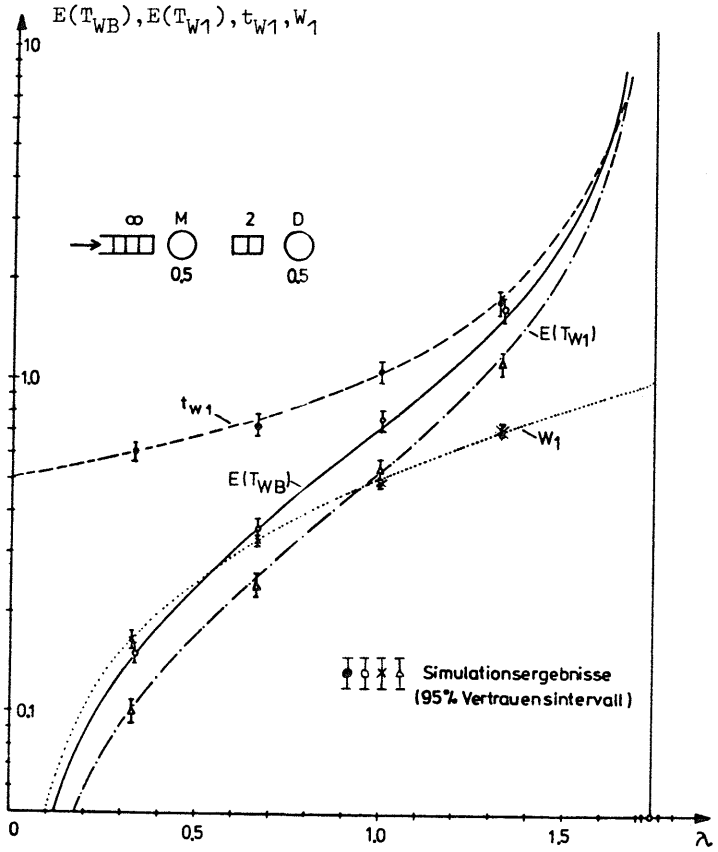


Bild 6.7 Mittlere Gesamtverzögerungszeit $E(T_{WB})$
mittlere Wartezeit aller Anforderungen in Stufe 1 $E(T_{W1})$
mittlere Wartezeit der Wartenden in Stufe 1 t_{W1}
und Wartewahrscheinlichkeit in Stufe 1 W_1
als Funktion der Ankunftsrate λ

Aus den Erwartungswerten $E(T_{WB}), E(T_{W1}), E(T_{W2}), E(T_B)$ ergeben sich in einfacher Weise die Wartebelastungen in beiden Stufen, und die Blockierbelastung in Stufe 1, wie auch die mittlere Gesamtzahl der sich im 2-stufigen System befindlichen Anforderungen.

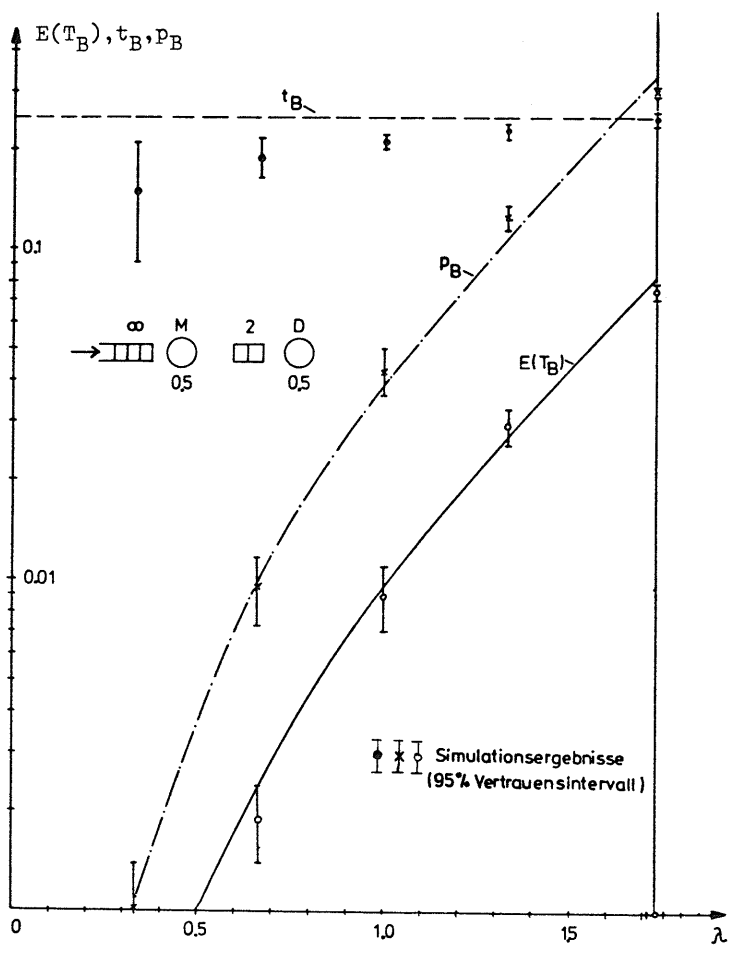


Bild 6.8 Blockier-Größen
Mittlere Blockierzeit aller Anforderungen $E(T_B)$
mittlere Blockierzeit der Blockierten t_B
Blockierwahrscheinlichkeit p_B

Als Wert für t_B wurde der in (5.26) approximativ für $\lambda = \lambda_{max}$ hergeleitete Wert herangezogen, der hier für $\lambda < \lambda_{max}$ eine Abschätzung nach oben darstellt. Dabei ist zu beachten, daß für $\lambda \ll \lambda_{max}$ die Blockierwahrscheinlichkeit p_B noch sehr klein ist.

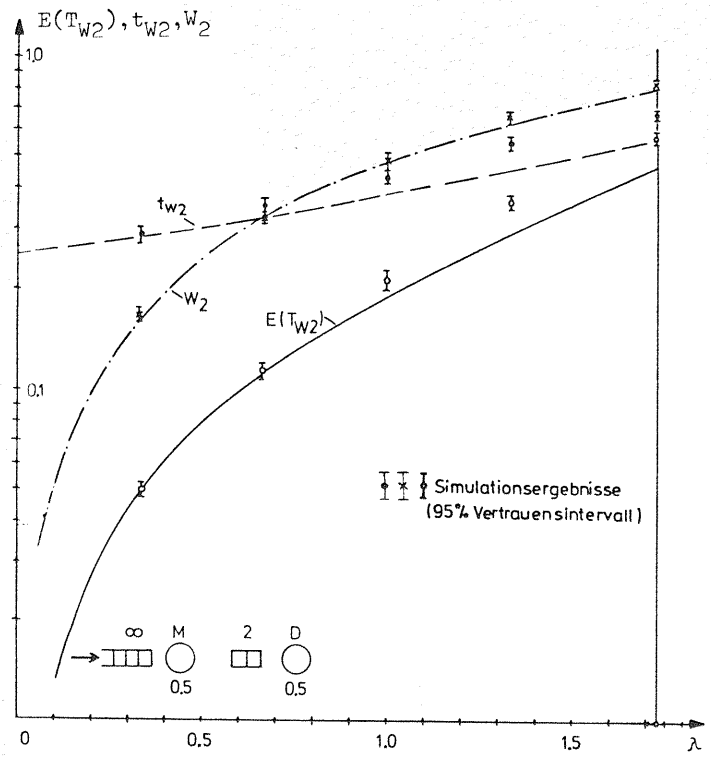


Bild 6.9 Größen im Wartespeicher der Stufe 2
Mittlere Wartezeit aller Anforderungen $E(T_{W2})$
mittlere Wartezeit der Wartenden t_{W2}
Wartewahrscheinlichkeit W_2

Tabelle 6.2 zeigt eine Zusammenstellung derjenigen Systeme, mit denen im folgenden versucht wird, einen Überblick über die Ergebnisse des Näherungsverfahrens und seiner Güte zu geben.

System Nr	VF_1-VF_2	h_1	h_2	s_2	
1	M-M	0.5	0.5	2	
2	M-E_2	0.5	0.5	2	
3	M-D	0.5	0.5	2	← vgl. auch Bilder 6.7-6.9
4	M-M	0.5	0.5	1	
5	M-M	0.5	0.5	5	
6	M-E_2	0.$\overline{3}$	0.$\overline{6}$	2	
7	M-E_2	0.$\overline{6}$	0.$\overline{3}$	2	
8	M-D	0.$\overline{3}$	0.$\overline{6}$	2	
9	M-D	0.$\overline{6}$	0.$\overline{3}$	1	
10	D-M	0.5	0.5	2	
11	D-E_2	0.5	0.5	2	
12	D-M	0.$\overline{3}$	0.$\overline{6}$	2	
13	D-M	0.5	0.5	1	
14	D-M	0.5	0.5	5	
15	D-E_2	0.$\overline{3}$	0.$\overline{6}$	2	
16	D-M	0.$\overline{6}$	0.$\overline{3}$	1	
17	D-E_2	0.$\overline{6}$	0.$\overline{3}$	2	
18	D-D	0.$\overline{3}$	0.$\overline{6}$	5	
(vgl.A2.4)	E_2-E_2	0.5	0.5	2	

Tabelle 6.2 Zusammenstellung der dargestellten Systeme

Wegen der Vielzahl der Ergebnisgrößen werden für diese Systeme nur die mittlere Gesamtverzögerungszeit $E(T_{WB})$ als wichtigste Größe des Systems bzw des Durchlaufs von Anforderungen dargestellt, wie auch $E(T_B)$ als Kenngröße für die Blockierung, die hier wesentlich den Durchlauf mitbeeinflussen kann.
In Bild 6.10-13 sind diese Größen für die Systeme 1 - 18 dargestellt und zwar für 4 verschiedene Ankunftsraten λ = 0.$\overline{3}$, 0.$\overline{6}$, 1.0 und 1.$\overline{3}$.

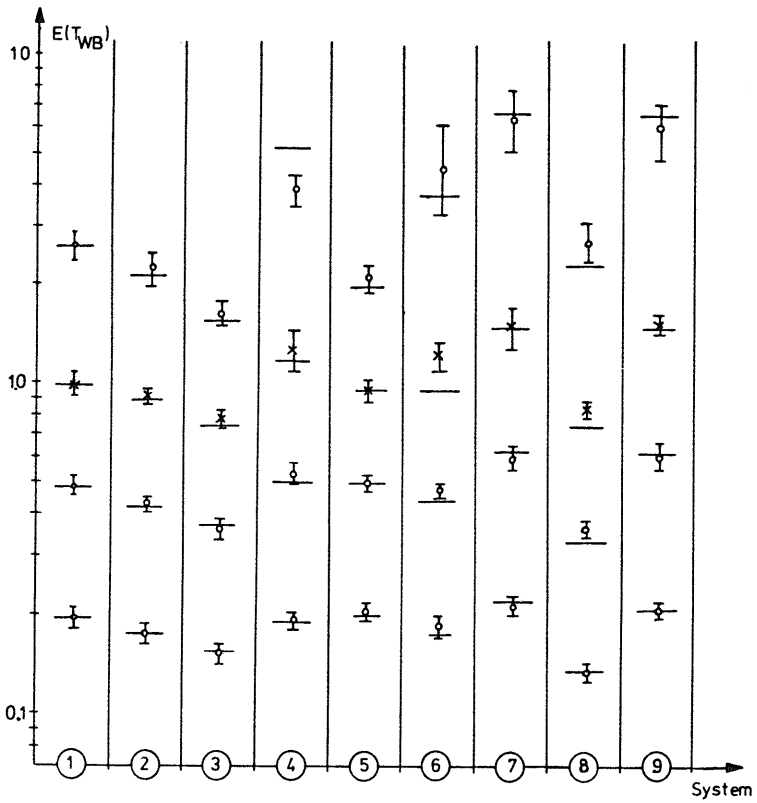


Bild 6.10 Mittlere Gesamtverzögerungszeiten $E(T_{WB})$ für Systeme mit VF_1=M (vgl. Tabelle 6.2). **Simulation bei 4 verschiedenen Ankunftsraten** ($\lambda = 0.\overline{3}$, $0.\overline{6}$, 1.0, $1.\overline{3}$) **Ergebnisse der Näherungsrechnung als horizontale Linien eingetragen.**

Wie aus Bild 6.10 hervorgeht, liegen die Ergebnisse des Näherungsverfahrens bei $E(T_{WB})$ für negativ-exponentiell verteilte Bedienungszeiten in Stufe 1 (VF_1=M) meist innerhalb der Vertrauensintervalle der Simulation. Lediglich bei System 6 war eine größere Abweichung (ca -20% bezogen auf den Mittelwert der Simulation) zu verzeichnen; die stärkere Abweichung von + 33% bei System 4 ist durch den asymptotischen Bereich zu erklären, in dem $E(T_{WB})$ für $\lambda = 1.\overline{3}$ ($\hat{=} \lambda/\lambda_{max} \approx 0.9$) bereits liegt.

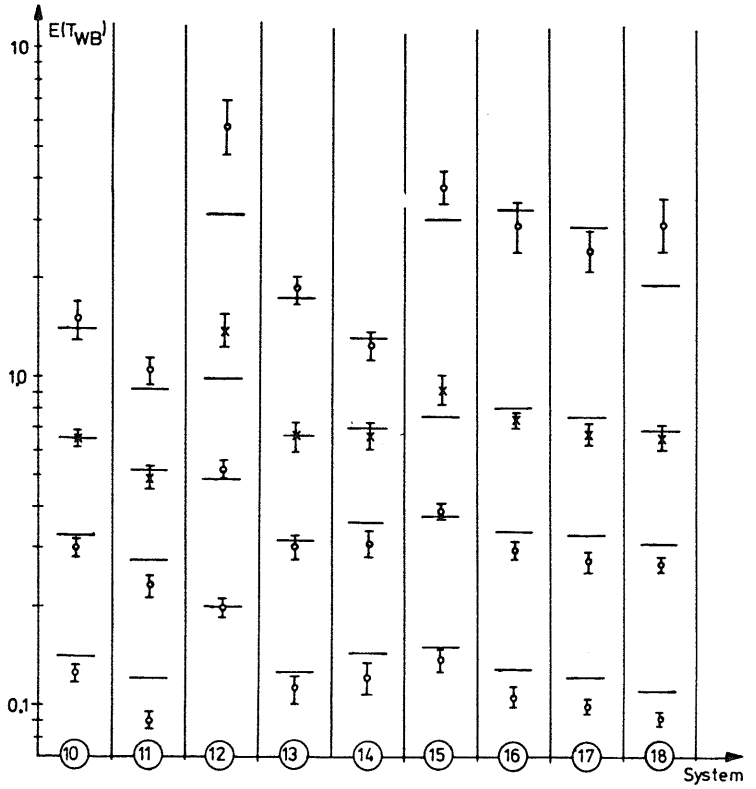


Bild 6.11 Mittlere Gesamtverzögerungszeiten $E(T_{WB})$ für Systeme mit VF_1=D (Legende vgl. Bild 6.10)

Für VF_1=D ist das Näherungsverfahren etwas weniger günstig als für VF_1=M (vgl. Bild 6.11). Berücksichtigt man, daß bei den höchsten Werten der Ankunftsrate der Systeme 12 und 18 $\lambda/\lambda_{max} \gtrsim 0.9$ gilt, so ergeben sich maximale Fehler von +35% bei System 11 und -40% bei 12.
Global kann für $E(T_{WB})$ gesagt werden, daß das Näherungsverfahren für $\lambda/\lambda_{max} \lesssim 0.8$ in den meisten Fällen gute Ergebnisse liefert, aber auch (abhängig von VF_1) Fehler bis ca 50% auftreten können.

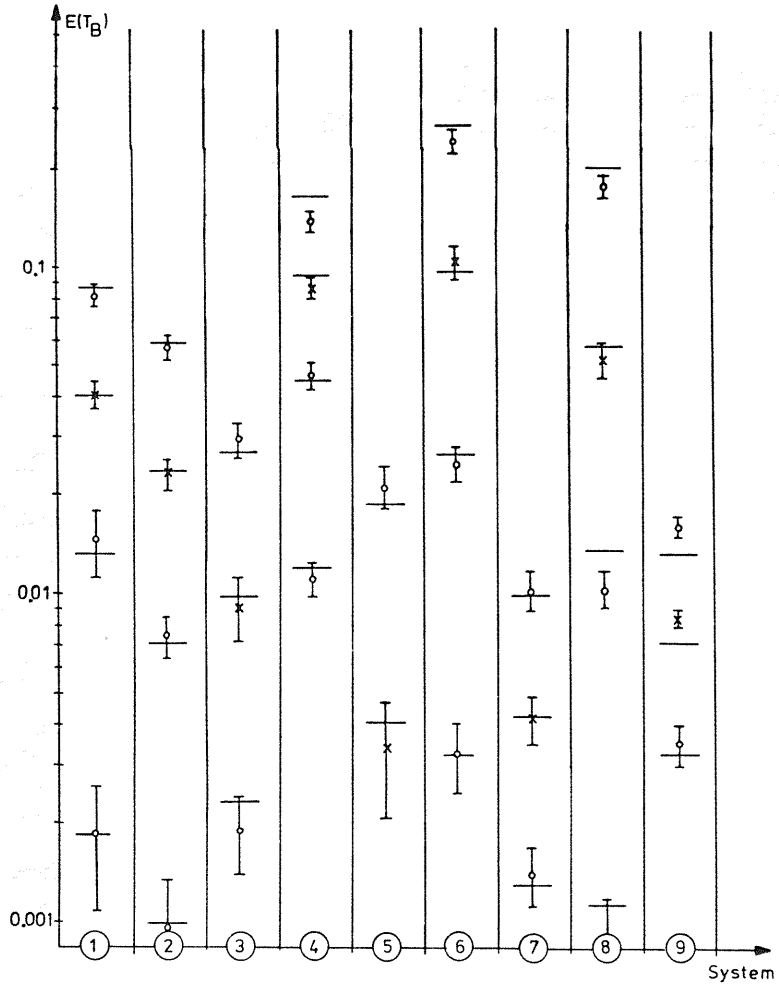


Bild 6.12 Mittlere Blockierzeiten aller Anforderungen $E(T_B)$ für Systeme mit VF_1=M (Legende vgl. Bild 6.10)

Die Bilder 6.12 und 6.13 für $E(T_B)$ sollen zeigen, inwiefern eine detailliertere Aussage (bezüglich der Blockierung) möglich ist, die hier prinzipiell eine entscheidende Rolle spielt.

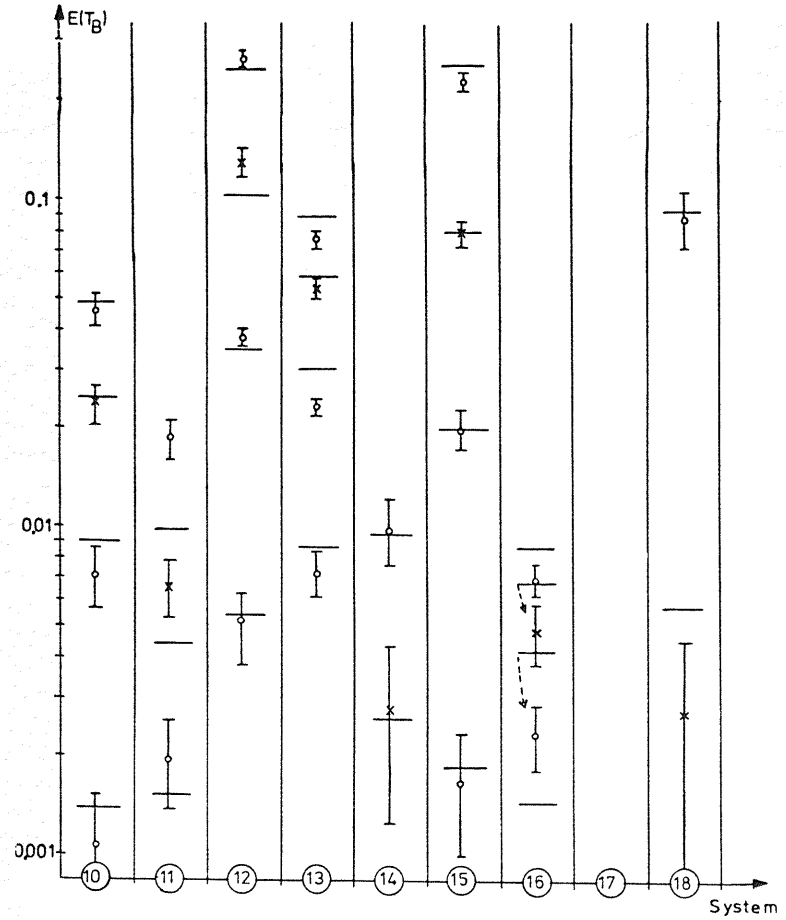


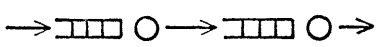
Bild 6.13 Mittlere Blockierzeiten aller Anforderungen $E(T_B)$ für Systeme mit VF_1=D (Legende vgl. Bild 6.10)

Bei System 11 mit relativ schwacher Blockierung auch bei maximaler Durchsatzrate tritt ein starker Fehler auf, der speziell auf einen zu niedrigen Näherungswert für t_B zurückzuführen ist. Bei System 16 wird $E(T_B)$ stark überschätzt, ist aber stets kleiner als 0.01, während bei System 17 $E(T_B)$ bei Simulation und Rechnung <0.001 waren.

Abschließend sei nochmals betont, daß es nicht das Ziel war, für einige wenige spezielle Systeme ein möglichst genaues Verfahren zu entwickeln, sondern ein Verfahren mit universellerer Anwendbarkeit auf beliebige derartige Systeme, bei dem jedoch größere Abweichungen in Kauf genommen werden mussten.

ZUSAMMENFASSUNG

Datenverarbeitungsanlagen und Nachrichtenvermittlungssysteme werden in immer größerer Zahl und Komplexität geplant,gefertigt und betrieben.Dabei ist es von großer Wichtigkeit,Voraussagen machen zu können über die "Betriebsgüte" eines solchen Systems,sei es bei der Planung zukünftiger oder der Modifikation bestehender Anlagen.
Da das Ablaufgeschehen in diesen Systemen durch spezielle stochastische (d.h. zufällige) Prozesse beschrieben werden kann,werden hierzu u.a. sog. Warteschlangenmodelle benutzt, die mit Methoden der Wahrscheinlichkeitstheorie berechnet werden.
Untersucht man den Verkehrsfluß in nachrichtenverarbeitenden, -vermittelnden und -übertragenden Systemen,wie auch in vielen anderen Bereichen (Fertigungsstraßen,Supermärkte usw.), erhält man dabei oft sog. seriell angeordnete Bedienungseinheiten (○),vor denen "Anforderungen" auf Warteplätzen (□) jeweils bis zum Beginn ihrer Bedienung warten können:

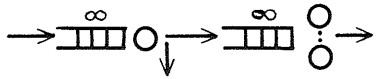


In der vorliegenden Arbeit wurden sog. 2-stufige Wartesysteme untersucht bei verschiedenen Voraussetzungen bezüglich der Systemstruktur (Zahl der Bedienungseinheiten und Warteplätze) sowie der Verkehre (Ankunfts- und Bedienungsprozesse).
Es wurden neue Ergebnisse abgeleitet wie Durchsatz,Wartezeiten usw. ,die für den planenden Systemingenieur bei der Dimensionierung von großem Nutzen sein können.

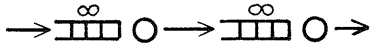
Im wesentlichen wurden Systeme untersucht mit je einer Bedienungseinheit in beiden Stufen,einem unendlich großen Wartespeicher in der 1.Stufe und sowohl endlich als auch unendlich großem Speicher (Puffer) zwischen beiden Bedienungseinheiten. Als Ankunftsprozeß in der 1.Stufe wurde ein Poissonprozeß angenommen,die Verteilungsfunktionen der Bedienungszeiten wurden meist nur durch ihre ersten beiden Momente charakterisiert.

Zunächst wurde in Kap.2 zur Motivation der Untersuchungen der nachfolgenden Kapitel ein Überblick gegeben über die in

der Literatur behandelten Wartesysteme in Serie und die wichtigsten dabei verwendeten analytischen Methoden mit ihren Ergebnissen dargestellt.
In Kap.3 wurde eine Untersuchung des Gesamtschicksals von Anforderungen durchgeführt in einem seriellen System mit unendlich großem Zwischenpuffer unter sog. rein Markoffschen Verkehrsannahmen.Dabei wurden zusätzlich Verzweigungen hinter der 1. Stufe zugelassen (Richtungsaufteilung),wie auch eine beliebige Zahl von Bedienungseinheiten in Stufe 2:

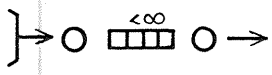


Durch das Verfolgen von Anforderungen durch das Gesamtsystem und die exakte Bestimmung der Abhängigkeiten der "Teilschicksale" einer Anforderung wurde u.a. die Verteilungsfunktion der Gesamtwartezeit exakt bestimmt.
In Kap.4 wurden ebenfalls Systeme mit unendlich großen Zwischenspeichern untersucht,d.h. Systeme ohne möglichen Rückstau in eine rückwärtige Stufe (Blockierung):

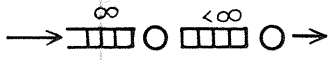


Der hier ungestörte Ausgangsprozeß der 1.Stufe ist in diesem allgemeinen Falle kein einfacher Erneuerungsprozeß,sondern ein Prozeß mit Abhängigkeiten zwischen aufeinanderfolgenden Ankunftsabständen.Mit Hilfe angebots- und varianzabhängiger Interpolation zwischen bekannten exakten und neuen approximativen Ergebnissen für Spezialfälle wurde eine Approximationsformel entwickelt für die mittlere Wartezeit aller Anforderungen in der 2.Stufe.Diese gilt für beliebig hypoexponentiell verteilte Bedienungszeiten in beiden Stufen.

Das in Kap.5 zugrundegelegte System bestand aus 2 seriell angeordneten Bedienungseinheiten mit endlich großem Zwischenpuffer und einem unendlich großen Reservoir mit Anforderungen:



Die wichtigste Verkehrsgröße dieses Systems ist der maximal erzielbare Durchsatz,der von der Größe des Zwischenpuffers und den beiden Bedienungszeit-Verteilungsfunktionen abhängt.Hierfür wurde eine einfache Approximationsformel hergeleitet,die nur die ersten beiden Momente der Bedienungszeitverteilungsfunktionen benötigt.
Schließlich wurde im letzten Kap.6 das zum System in Kap.5 zugehörige verlustfreie System mit Poissonankünften in Stufe 1 und hypoexponentiellen VF der Bedienungszeiten bei nicht gesättigter 1.Stufe untersucht:



Dabei wurde das System trotz Rückstaumöglichkeit in 2 quasi-unabhängige Teilsysteme aufgeschnitten und der Rückstau approximativ durch eine vergrößerte Gesamtbelegungsdauer in Stufe 1 berücksichtigt.Neben den Verkehrsgrößen für die individuellen Stufen konnte damit z.B. auch die mittlere Durchlaufszeit für das ganze System angegeben werden.
Alle Ergebnisse der Arbeit,soweit sie approximative Lösungen darstellen,wurden durch ausführliche Simulationen abgesichert.

ANHANG

A1 Einige benutzte Verteilungsfunktionen

A1.1 Allgemeines

Die Verteilungsfunktion (VF) einer (hier kontinuierlichen) Zufallsvariablen (z.B. Bedienungszeit T_H) gibt an,mit welcher Wahrscheinlichkeit diese Zufallsvariable kleiner oder höchstens gleich einer Zeit t ist:

$$P(T_H \leqslant t) \overset{def}{=} H(\leqslant t) \tag{A1.1}$$

Oft betrachtet man auch die komplementäre VF

$$P(T_H > t) \overset{def}{=} H(>t) = 1 - H(\leqslant t) \tag{A1.2}$$

Die zugehörige Dichtefunktion ist

$$f_H(t) = \frac{dH(\leqslant t)}{dt} \tag{A1.3}$$

Der Erwartungswert der Zufallsvariablen beträgt:

$$E(T_H) \overset{def}{=} h = \int_0^\infty t \cdot f_H(t)dt = \int_0^1 t \cdot dH(\leqslant t) \tag{A1.4}$$

Das j. Moment der VF der Zufallsvariablen T_H bezüglich einer Größe a ist definiert als der Erwartungswert

$$E([T_H-a]^j)$$

Bei a=0 spricht man von den gewöhnlichen Momenten,die hier mit m_j abgekürzt seien,bei $a=E(T_H)$ von zentralen Momenten oder Momenten bezüglich des Mittelwerts.
Das 2. zentrale Moment einer VF wird Varianz oder Streuungsquadrat genannt

$$\sigma^2 \overset{def}{=} E([T_H-E(T_H)]^2) = \int_0^\infty [t-E(T_H)]^2 \cdot f_H(t)dt \overset{def}{=} Var(T_H) \tag{A1.5}$$

und kann auch als

$$\sigma^2 = E(T_H^2) - E(T_H)^2 = m_2 - m_1^2 \tag{A1.6}$$

dargestellt werden. σ selbst wird als Standardabweichung bezeichnet.
Das charakteristische Verhältnis

$$C \overset{def}{=} \frac{\sigma}{E(T_H)} \tag{A1.7}$$

stellt den Varianzkoeffizienten dar.

Der Varianzkoeffizient einer negativ-exponentiellen VF ist 1,VF mit C<1 sind vom Typ hypoexponentiell,während VF mit C>1 als hyperexponentiell bezüglich der Varianz gelten.
Als (einzige hier angegebene) VF einer diskreten Zufallsvariablen sei die sog. Poisson-VF genannt.

$$p(i,t_o) = \frac{(\lambda t_o)^i}{i!} \cdot e^{-\lambda t_o} \qquad (A1.8)$$

Sie gibt an,mit welcher Wahrscheinlichkeit bei einem reinen Ge-Geburtsprozeß/Sterbeprozeß (Geburtsabstände/Sterbeabstände negativ--exponentiell verteilt mit Mittelwert $1/\lambda$) genau i Geburten/Sterbefälle in ein festes Zeitintervall der Länge t_o fallen.

Damit gilt für die Wahrscheinlichkeit q_i,daß im Wartesystem M|G|1 während einer beliebigen nach G verteilten,zufälligen Bedienungszeit genau i Anforderungen ankommen

$$q_i = \int_0^\infty \frac{(\lambda t)^i}{i!} \cdot e^{-\lambda t} \cdot f_H(t)dt \qquad (A1.9)$$

Diese,sagen wir einmal Ankunftswahrscheinlichkeiten q_i,werden in Kapitel 6 benötigt,deshalb sind sie bei den verschiedenen VF mit angegeben.
Da $H(\leqslant t) \equiv 0$ für t<0,gilt für alle angegebenen VF $0 \leqslant t < \infty$ und als Parameter $k \geqslant 1, j \geqslant 1, i \geqslant 0$. Außerdem wurde $A = \lambda \cdot h$ als Abkürzung verwendet.

In A1.2-A1.5 werden einige wichtige Standard-VF beschrieben,während in A1.6-A1.7 weitere VF angegeben sind,die zur Untersuchung des Einflusses des 3.und höherer Momente von VF verwendet wurden (vgl A2).

A1.2 Negativ-exponentielle VF (M)

Die negativ-exponentielle VF gilt als wichtigste VF und besitzt keinerlei Gedächtnis (vgl. 1.2.1).

$$\left.\begin{array}{lll} H(\leqslant t) = 1 - e^{-\varepsilon t} & & f_H(t) = \varepsilon \cdot e^{-\varepsilon t} \\ E(T_H) = h = \frac{1}{\varepsilon} & \sigma_H^2 = \left(\frac{1}{\varepsilon}\right)^2 & C_H = 1 \\ m_j = j! \cdot h^j & & q_i = \frac{A^i}{(1+A)^{i+1}} \end{array}\right\} \quad (A1.10)$$

A1.3 Erlang-k-VF (E_k)

Eine Erlang-k-verteilte Zufallsvariable kann als Summe von k negativ-exponentiellen Zufallsvariablen dargestellt werden.

$$\left.\begin{array}{lll} H(\leqslant t) = 1 - e^{-\varepsilon t} \cdot \sum_{\nu=0}^{k-1} \frac{(\varepsilon t)^\nu}{\nu!} & & f_H(t) = \varepsilon \cdot e^{-\varepsilon t} \cdot \frac{(\varepsilon \cdot t)^{k-1}}{(k-1)!} \\ E(T_H) = h = \frac{k}{\varepsilon} & \sigma_H^2 = \frac{k}{\varepsilon^2} & C_H = \frac{1}{\sqrt{k}} \\ m_j = \frac{(k-1+j)!}{(k-1)!\ k^j} \cdot h^j & & q_i = \binom{i+k-1}{i} \left(\frac{A}{A+k}\right)^i \cdot \left(\frac{k}{A+k}\right)^k \end{array}\right\} \quad (A1.11)$$

A1.4 Konstante VF (D)

Sie ist eine entartete Verteilung mit Varianz 0.

$$\left.\begin{array}{lll} H(\leqslant t) = \begin{cases} 0 & t<h \\ 1 & t \geqslant h \end{cases} & & f_H(t) = \delta(t-h) \\ E(T_H) = h & \sigma_H^2 = 0 & C_H = 0 \\ m_j = h^j & & q_i = \frac{A^i}{i!} \cdot e^{-A} \end{array}\right\} \quad (A1.12)$$

A1.5 Hyperexponentielle VF (H_k)

Eine hyperexponentielle VF k.Ordnung ist dadurch beschreibbar,daß mit einer Wahrscheinlichkeit p_ν (ν=1..k) eine negativ-exponentielle Phase mit individuellem Mittelwert $1/\varepsilon_\nu$ durchlaufen wird.

$$H(\leqslant t) = 1 - \sum_{\nu=1}^{k} p_\nu e^{-\varepsilon_\nu t} \qquad f_H(t) = \sum_{\nu=1}^{k} p_\nu \cdot \varepsilon_\nu \cdot e^{-\varepsilon_\nu t}$$

$$E(T_H) = h = \sum_{\nu=1}^{k} p_\nu \cdot \frac{1}{\varepsilon_\nu} \qquad \sigma_H^2 = 2\left(\frac{p_1}{\varepsilon_1^2} + \cdots + \frac{p_k}{\varepsilon_k^2}\right) - \left(\frac{p_1}{\varepsilon_1} + \cdots + \frac{p_k}{\varepsilon_k}\right)^2 \qquad C_H > 1$$

$$m_j = j! \cdot \sum_{\nu=1}^{k} p_\nu \cdot \left(\frac{1}{\varepsilon_\nu}\right)^j \qquad q_i = \sum_{\nu=1}^{k} p_\nu \cdot \frac{A_\nu^i}{(1+A_\nu)^{i+1}} \qquad (A1.13)$$

$$\text{mit } A_\nu = \frac{\lambda}{\varepsilon_\nu}$$

Da eine hyperexponentielle VF bei vorgeschriebener Varianz weitere Freiheitsgrade besitzt,mußten zur Realisierung fester Varianzen noch Zusatzbedingungen angenommen werden.

Diese waren

$$k = 2 \quad \text{und} \quad p_1 \cdot \frac{1}{\varepsilon_1} = p_2 \cdot \frac{1}{\varepsilon_2}$$

bei der speziellen in Kapitel 5 verwendeten hyperexponentiellen VF (H_2). Mit $p_1=p, p_2=1-p$ folgt aus (A1.13):

$$C_H = \sqrt{\frac{1}{2}\left(\frac{1}{p} + \frac{1}{(1-p)}\right) - 1}$$

Parameterwahl bei vorgegebenem Mittelwert h und Varianzkoeffizient C_H:

$$p = \frac{1 \overset{(+)}{-} \sqrt{1 - \frac{2}{1+C_H^2}}}{2} \qquad \varepsilon_1 = \frac{2p}{h} \qquad \varepsilon_2 = \frac{2(1-p)}{h} \qquad \text{(A1.14)}$$

A1.6 "Verschobene" negativ-exponentielle VF ("DM")

Derart verteilte Zufallsvariablen erhält man durch Addition einer festen Zeit h_1 zu negativ-exponentiell verteilten Zufallsvariablen mit Mittelwert $h_2=1/\varepsilon_2$. Damit können -im Gegensatz zu einer E_k-VF - alle Varianzkoeffizienten $0 \leqslant C_H \leqslant 1$ realisiert werden.

$$H(<t) = \begin{cases} 0 & t<h_1 \\ 1-e^{-\varepsilon_2(t-h_1)} & t \geqslant h_1 \end{cases} \qquad f_H(t) = \begin{cases} 0 & t<h_1 \\ \varepsilon_2 \cdot e^{-\varepsilon_2(t-h_1)} & t \geqslant h_1 \end{cases}$$

$$E(T_H) = h_1 + h_2 \qquad \sigma_H^2 = h_2^2 \qquad C_H = \frac{h_2}{h_1 + h_2} \qquad \text{(A1.15)}$$

$$m_j = \varepsilon_2 e^{\varepsilon_2 h_1} \cdot \int_{h_1}^{\infty} t^j e^{-\varepsilon_2 t}\, dt \qquad q_i = \varepsilon_2 e^{\varepsilon_2 h_1} \cdot \frac{\lambda^i}{i!} \cdot \int_{h_1}^{\infty} t^i \cdot e^{-(\lambda+\varepsilon_2)t}\, dt$$

(Integrale rekursiv auswertbar)

Diese VF wurde in den Kapiteln 4-6 verwendet (vgl.auch A2).

A1.7 Mehrpunkt-VF mit k Punkten (P_k)

Dies ist eine Verallgemeinerung einer konstanten VF, bei der eine Zufallsvariable mit der Wahrscheinlichkeit p_ν einen bestimmten Wert h_ν annimmt.

$$H(\leqslant t) = \sum_{\nu=1}^{k} p_\nu \cdot s(t-h_\nu) \qquad f_H(t) = \sum_{\nu=1}^{k} p_\nu \cdot \delta(t-h_\nu) \qquad \text{(A1.16)}$$

$$E(T_H) = \sum_{\nu=1}^{k} p_\nu \cdot h_\nu \qquad \sigma_H^2 = \sum_{\nu=1}^{k} p_\nu (h_\nu - h)^2 \qquad C_H = \sqrt{\sum_{\nu=1}^{k} p_\nu \left(\frac{h_\nu}{h} - 1\right)^2}$$

$$m_j = \sum_{\nu=1}^{k} p_\nu \cdot h_\nu^j \qquad q_i = \sum_{\nu=1}^{k} p_\nu \cdot \frac{(\lambda h_\nu)^i}{i!} \cdot e^{-\lambda h_\nu}$$

Es wurden 2 spezielle Mehrpunkt-VF untersucht:

a) $k=2, p_1=p, p_2=1-p$, $h_2=0$ (Spezielle 2-Pkt-VF, P_2)

$$E(T_H) = p \cdot h_1 \qquad \sigma_H^2 = p(1-p) \cdot h_1^2 \qquad C_H = \sqrt{\frac{1-p}{p}}$$

$$m_j = p \cdot h_1^j \qquad q_i = \begin{cases} p e^{-\lambda h_1} + (1-p) & i=0 \\ p \cdot \frac{(\lambda h_1)^i}{i!} \cdot e^{-\lambda h_1} & i>0 \end{cases} \qquad \text{(A1.17)}$$

Parameterwahl bei vorgegebenem Mittelwert $E(T_H)$ und Varianzkoeffizient C_H:

$$p = \frac{1}{(1+C_H^2)} \qquad h_1 = (1+C_H^2) \cdot E(T_H)$$

b) $k \to \infty, p_\nu = (1-p) \cdot p^{\nu-1}$ $(\nu>0)$, $h_\nu = \nu \cdot T$ (Geometrische ∞-Pkt-VF, P_∞)
($T \hat{=}$ "Taktzeit")

$$E(T_H) = \frac{T}{(1-p)} \qquad \sigma_H^2 = p \cdot \frac{T^2}{(1-p)^2} \qquad C_H = \sqrt{p} \qquad \text{(A1.18)}$$

Parameterwahl bei vorgegebenem Mittelwert $E(T_H)$ und Varianzkoeffizient C_H:

$$p = C_H^2 \qquad T = (1-C_H^2) \cdot E(T_H)$$

A2 Untersuchung des Einflusses höherer Momente

A2.1 Allgemeines

Bei der Konzipierung der Näherungsverfahren in Kapitel 4,5 und 6 wurde zur universelleren Anwendbarkeit von Anfang an die VF nur durch Mittelwert und Varianz berücksichtigt. Dabei lag die Motivation zugrunde, daß

a) **die mittleren Wartezeiten im System M\|G\|1 exakt nur von den ersten beiden Momenten der Bedienungszeit-VF abhängen**
b) bei Überlaufsystemen z.B. die Berücksichtigung von nur 2 Momenten bereits recht brauchbare Ergebnisse liefert
c) in der Praxis häufig nur diese beiden Parameter bekannt sind

bzw. abgeschätzt werden können.
Neben den untersuchten VF M, E_2 und D wurden die im Anhang A1 aufgezählten VF untersucht und die Näherungsverfahren auf entsprechende Systeme angewendet. Bild A2.1 zeigt als Beispiel hierfür verschiedene VF mit $C_H^2 = 0.5$.

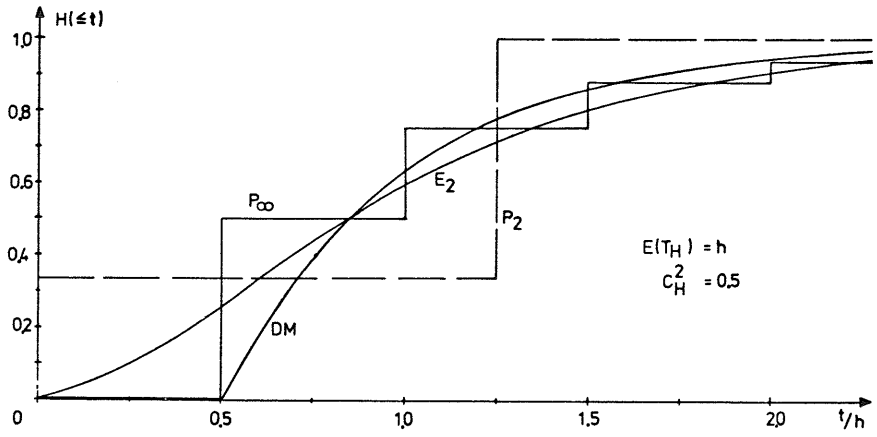


Bild A2.1 Vergleich verwendeter VF-Typen
(Parameterberechnung nach A1)

Diese VF besitzen gleiches 1. und 2. Moment aber verschiedene höhere Momente, die in A1 angegeben wurden.
Neben diesen VF erfolgt auch ein Vergleich bei $C_H^2 = 1$, nämlich zwischen
- negativ-exponentieller VF (M)
  mit den Momenten $m_j/h^j = j!$ und
- spezieller 2-Punkt-VF (P_2)
  mit den Momenten $m_j/h^j = 2^{j-1}$ $(j>0)$

also sehr verschiedenen höheren Momenten.
Die nachfolgenden Untersuchungsergebnisse zeigen in großem Ganzen die Berechtigung der Näherungsannahme der Beschreibung durch nur 2 Momente. Dabei muß allerdings gesagt werden, daß VF denkbar sind, bei denen diese Verfahren schlechte Ergebnisse liefern. Hierzu müssen jedoch ganz spezielle Bedingungen erfüllt sein.
Betrachtet man z.B. die extreme VF-Kombination D-P_2 mit $h_{12}=(1+C_{H2}^2)\cdot h_2$ und $h_{22}=0$ als Parameter der P_2-VF in Stufe 2, so ergibt sich für

$$h_1 \geq h_{12} \qquad \frac{h_1}{h_2} \geq (1+C_{H2}^2)$$

im realen System kein Warten in Stufe 2 und keine Blockierung, dies ergibt sich bei den Näherungsverfahren nur bei $C_{H2}^2=0$ (VF_2=D). Bei $C_{H2}^2=0.5$ beispielsweise ergeben die Verfahren stets endlich große Warte- und/oder Blockierzeiten, die realen Systeme jedoch nur für $h_1/h_2<1.5$.
Allgemein kann gesagt werden, daß die ganz speziellen VF-Kombinationen D-VF_2^{**} und VF_1^{**}-D für die Verfahren nicht geeignet sind, falls
- VF_2^{**} einen Varianzkoeffizienten $C_{H2}>0$ und einen Maximalwert der Bedienungszeit $h_{2max}<h_1$
- VF_1^{**} einen Varianzkoeffizienten $C_{H1}>0$ und einen Minimalwert der Bedienungszeit $h_{1min}\geq h_2$

besitzt, was aber eine unwesentliche Einschränkung bedeutet.

A2.2 Ergänzungen zu Kapitel 4

Für den Fall 2-stufiger Single-Server Systeme ohne Blockierung (Kapitel 4) seien hier verschiedene Beispiele angegeben.

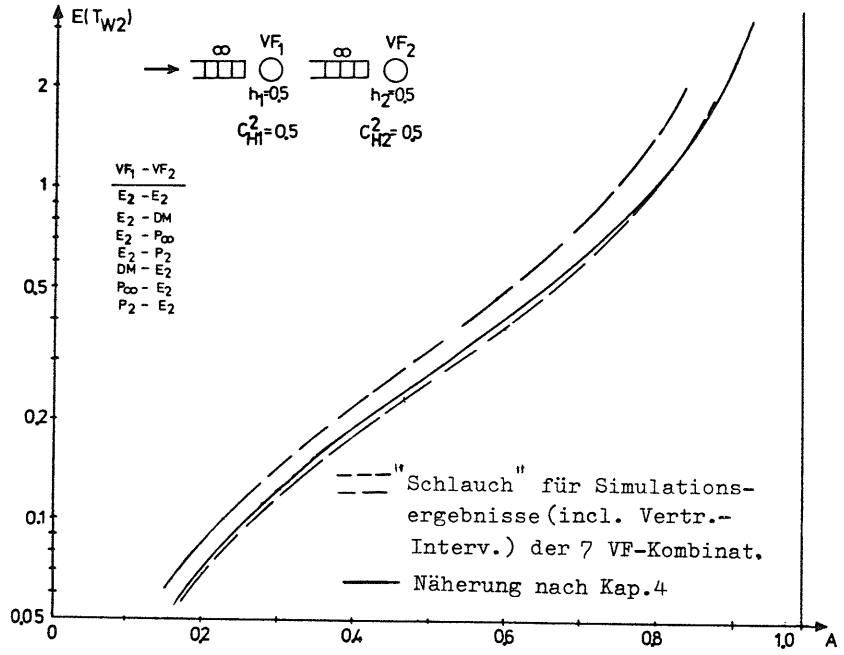


Bild A2.2 Mittlere Wartezeiten in Stufe 2 für verschiedene VF-Kombinationen

Bild A2.2 zeigt den durch Simulation gewonnenen Verlauf der mittleren Wartezeiten in Stufe 2 $E(T_{W2})$ über dem Angebot $A_1=A_2=A$ für 7 verschiedene VF-Kombinationen mit $C^2_{H1}=C^2_{H2}=0.5$ .Dabei ist nur ein "Schlauch" angegeben,innerhalb dessen alle Ergebnisse der Simulation für A=0.2,0.4,0.6,0.8 inclusive Vertrauensintervalle lagen.Die durchgezogene Kurve ist die Näherung nach Kapitel 4, die hier lediglich für $A\gtrsim 0.85$ außerhalb des Schlauches liegt. Insgesamt kann gesagt werden,daß $E(T_{W2})$ eine gewisse Robustheit gegenüber dem speziellen Typ der VF zeigt.
Bild A2.3 zeigt 2 Systeme,bei denen bei fester Ankunftsrate λ die Varianz der VF DM variiert wurde und die durch Vertauschung der Reihenfolge der Stufen auseinander hervorgehen.

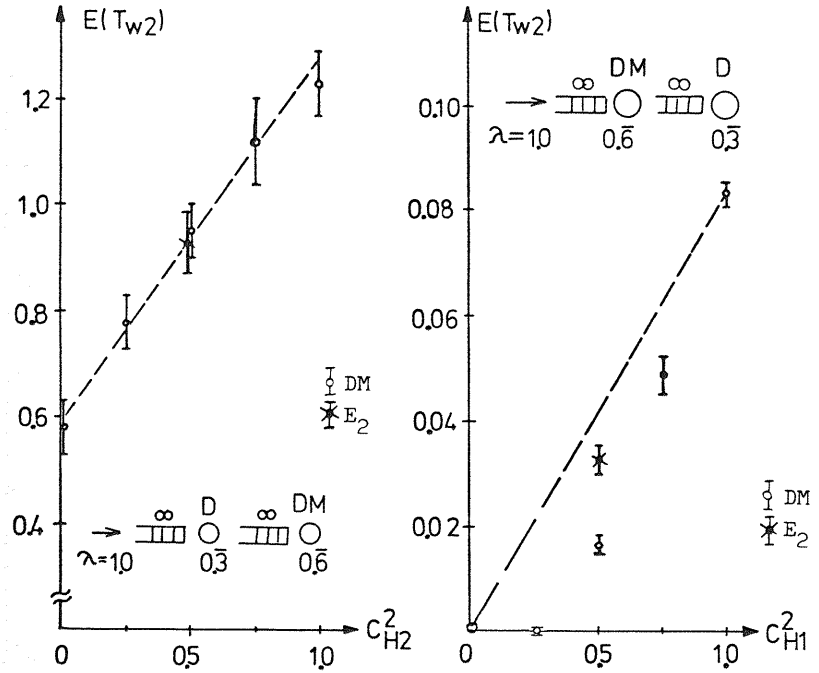


Bild A2.3 Mittlere Wartezeiten in Stufe 2
a) bei D-DM b) bei DM-D

Zum Vergleich sind die Simulationsergebnisse für E_2 ($C^2_H=0.5$) mit eingetragen.Dabei soll Bild A2.3b) als Beispiel für eine in A2.1 erwähnte spezielle VF-Kombination VF**-D gelten:

VF_1^{**} besitzt den Minimalwert $h_{11}=(1-C_{H1})\cdot E(T_{H1})$ (konstanter Anteil), so daß für $h_{11}\geq h_2$ nie eine Anforderung in Stufe 2 warten muß,also hier für

$$1-C_{H1}\geq \frac{h_2}{h_1}=0.5 \quad \rightsquigarrow \quad C^2_{H1}\leq 0.25$$

Hieraus erklärt sich die im Vergleich zur 2-Momenten-Approximation geringere Wartezeit für $C^2_{H1}=0.5$,während für E_2 die Näherung noch brauchbar ist.
Dabei sollte man nicht übersehen,daß bei dieser Ankunftsrate $\lambda=1.0$ ($A_{max}=0.\bar{6}$) dies jedoch bei $E(T_{W2})/E(T_{W1})\approx 1/30$ auftritt.

A2.3 Ergänzungen zu Kapitel 5

Bei 2 BE in Serie mit endlichem Zwischenspeicher wurde in Kapitel 5 der maximale Durchsatz approximativ bestimmt.Bild A2.4a,b zeigt beispielhaft einige Simulationsergebnisse für $s_2=2$ und $C^2_{H1}+C^2_{H2}=1.0$ bzw. 1.5 für verschiedenste VF-Kombinationen im Vergleich mit der Näherung.

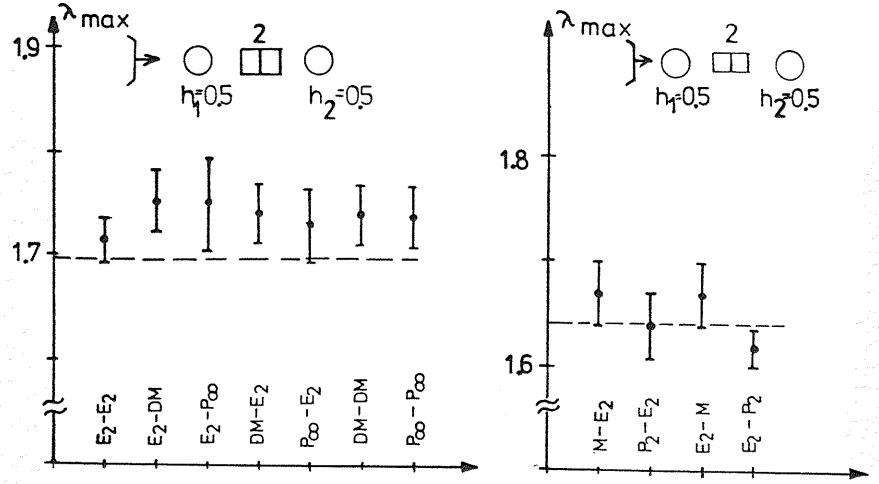


Bild A2.4 Maximaler Durchsatz
a) bei $C^2_{H1}+C^2_{H2}=1.0$ b) bei $C^2_{H1}+C^2_{H2}=1.5$

Auch hier ist eine gewisse Robustheit der Ergebnisse bei gleichem 1. und 2.Moment gegenüber den höheren Momenten festzustellen.

A2.4 Ergänzungen zu Kapitel 6

Hierzu zeigt Bild A2.5 einen Schlauch mit durchgezogener Näherung nach Kapitel 6 für $E(T_{WB})$, in dem alle Simulationsergebnisse der 5 angegebenen VF-Kombinationen mit $C_H^2=0.5$ lagen, inclusive Vertrauensintervalle.

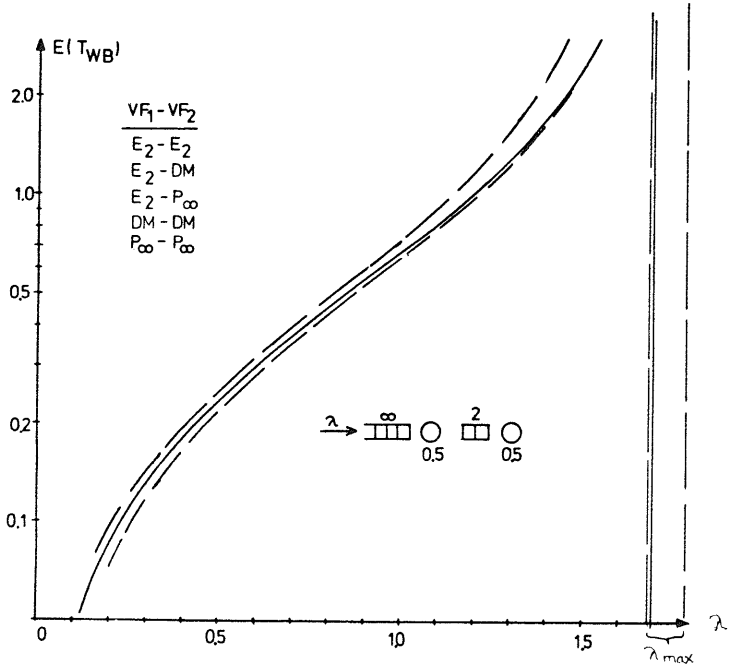


Bild A2.5 Mittlere Gesamtverzögerungszeit für verschiedene VF-Kombinationen mit $C_H^2=0.5$ und $h_1=h_2=0.5$

Bei $E(T_{W1})$ ergab sich praktisch derselbe prinzipielle Verlauf des Schlauches.
Betrachtet man die einzelnen Komponenten von $E(T_{WB})$ in Bild A2.6, so erkennt man auch bei $E(T_{W2})$ und $E(T_B)$ die relativ gute Brauchbarkeit des Verfahrens.
Lediglich bei den Wahrscheinlichkeiten (W_2 und p_B) wichen die Ergebnisse der speziellen VF-Kombination P_∞-P_∞ etwas stärker von der Näherung ab.

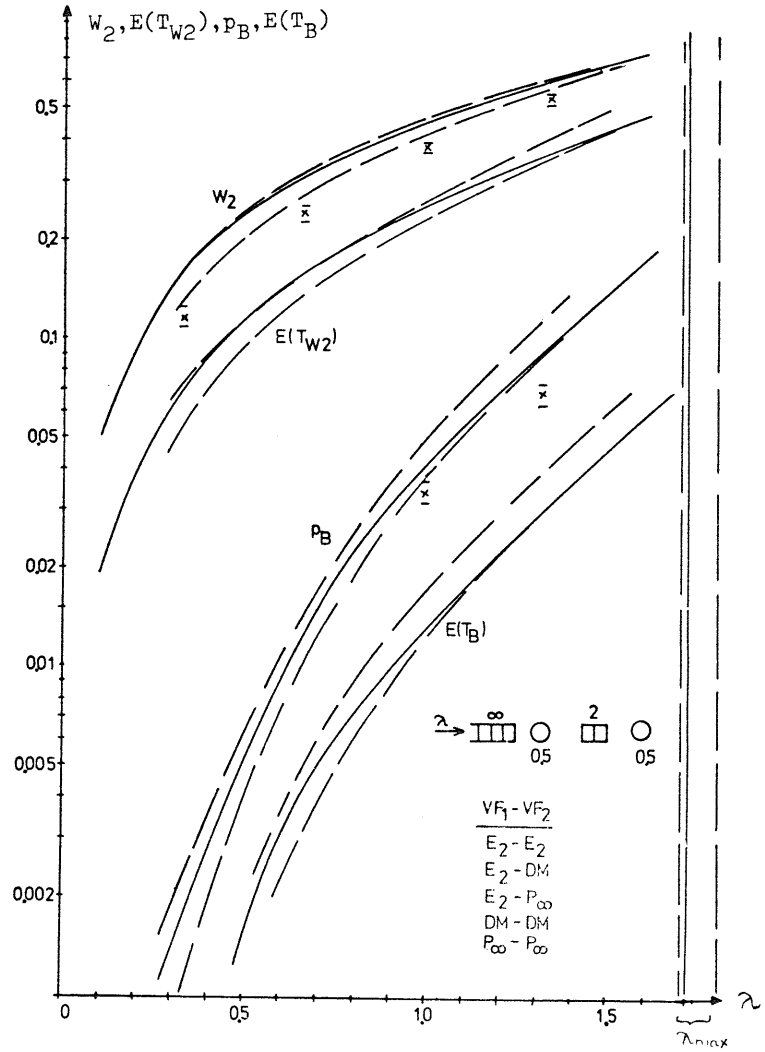


Bild A2.6 W_2, $E(T_{W2})$, p_B und $E(T_B)$ bei verschiedenen VF-Kombinationen (Ergebnisse für P_∞-P_∞ bei Wahrscheinlichkeiten teilweise separat angegeben $\bar{x}$)

Dies ist darauf zurückzuführen, daß W_2 und p_B (ebenfalls wie t_{W2} und t_B) stärker vom speziellen Typ (und damit von den höheren Momenten) der VF abhängen, da sie eine detailliertere Aussage über die VF von T_{W2} und T_B darstellen, als lediglich deren Erwartungswerte.

A3 Ergänzende Diagramme zu Kapitel 5

Da sich die maximale Durchsatzrate λ_{max} bei Vertauschung der mittleren Bedienungszeiten h_1 und h_2 praktisch nur wenig ändert (nach Satz 5.3), kann λ_{max} als Funktion des Verhältnisses

$$\frac{h_{min}}{h_{max}} = \frac{\min(h_1,h_2)}{\max(h_1,h_2)}$$ dargestellt werden.

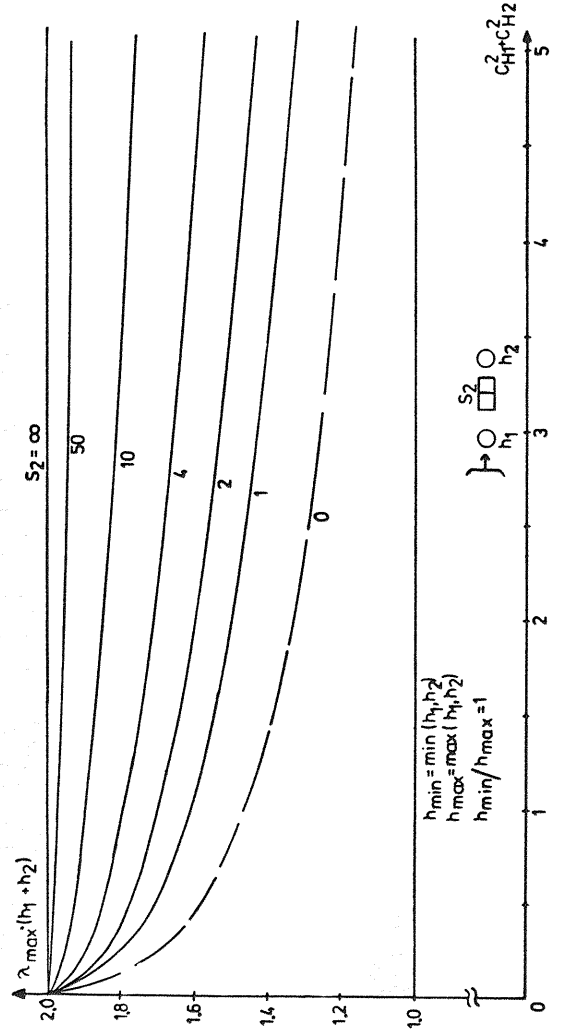


Bild A3.1 Maximaler Durchsatz ($h_1=h_2$)

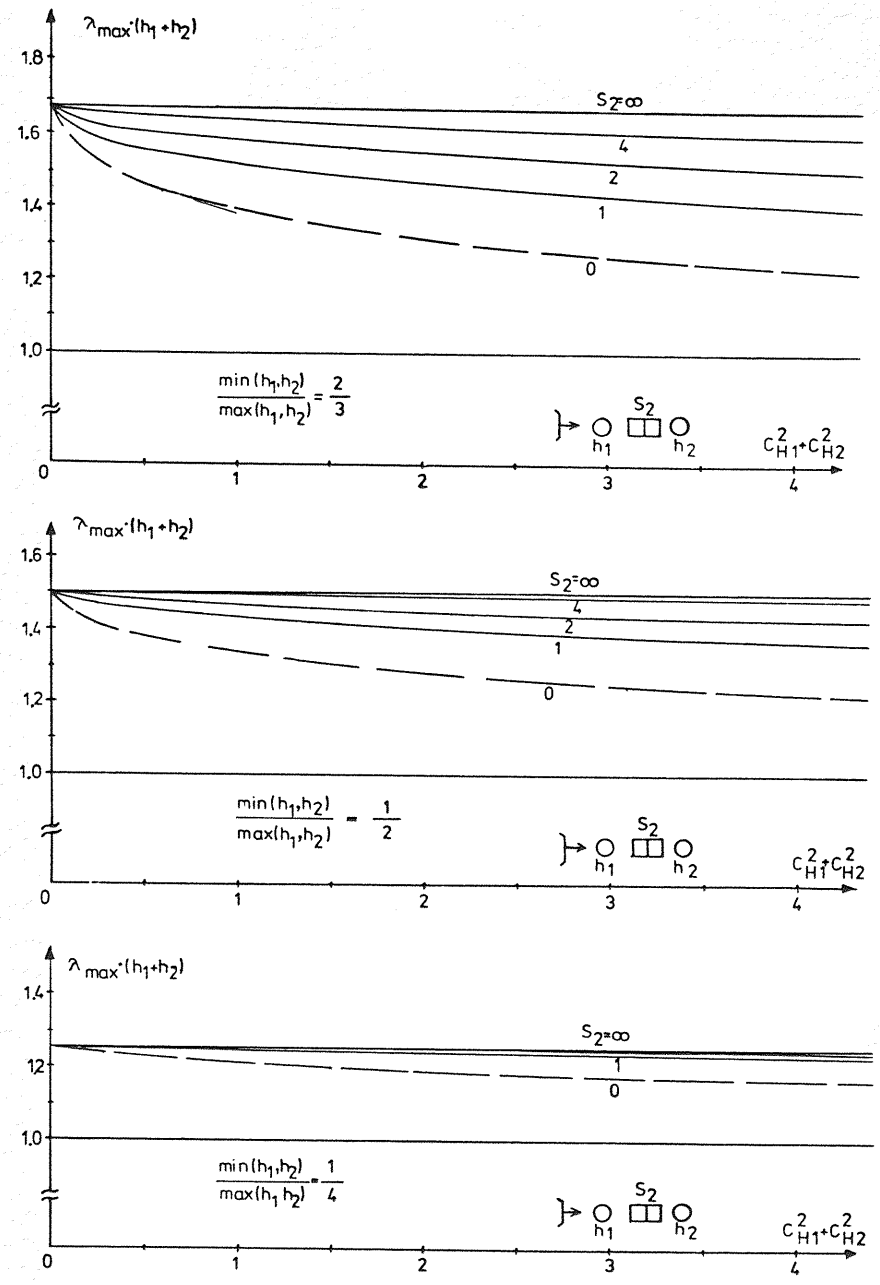


Bild A3.2a,b,c Maximaler Durchsatz ($h_1\neq h_2$)