Institut für Nachrichtenvermittlung und Datenverarbeitung
Universität Stuttgart
Prof. Dr.-Ing. P. Kühn

36. Bericht über verkehrstheoretische Arbeiten

Überlastprobleme in rechnergesteuerten Fernsprechvermittlungssystemen — Modellbildung und Analyse

von
Phuoc TRAN-GIA

Institute of Switching and Data Technics
University of Stuttgart
Prof. Dr.-Ing. P. Kühn

36th Report on Studies in Congestion Theory

Modelling Overload Control in SPC Switching Systems

by
Phuoc TRAN-GIA

1982

Druck: E. Kurz & Co.

ISBN 3-922403-46-8

ABSTRACT

During the past decade, a number of digital, stored program controlled (SPC) switching systems have been developed and introduced which replace the electromechanical switching system generation. While these systems have advantages in hard- and software design methodology and have an increasing number of new system features as well as customer facilities, the overload phenomenon and its influence to the system performance has become more critical and complex. In order to guarantee a proper system performance the sensitivity of switching systems against overload must be taken into account in the design and development phase as well as in the post-cutomer phase of a switching system.

This report deals with modelling and analysis aspects of the overload phenomenon and overload control strategies in stored program controlled telephone switching systems.

CHAPTER 1 INTRODUCTION

General aspects of the overload phenomenon and the problem of overload control in telephone switching systems are considered in this chapter. Also an outline of the objectives of the report is given.

CHAPTER 2 FUNCTIONS AND CONTROL STRUCTURES OF STORED PROGRAM CONTROLLED SWITCHING SYSTEMS

The first part of this chapter deals with basic functions of a telephone switching system, where control structures containing autonomous switching processes are introduced and briefly described. Subsequently, control principles and architectures of modern SPC systems, e.g. centralized, decentralized and distributed control structures, are discussed. For several system architectures the principles of interprocess- and interprocessor-communications are outlined. In the last part of this chapter the three load levels, call, subprocess for call handling and subcall traffic levels are introduced which will be used to characterize traffic streams in switching systems.

CHAPTER 3 OVERLOAD PROBLEMS IN SPC SWITCHING SYSTEMS

Definitions of overload and of overload control are given in the first part of this chapter. In section 3.2 overload sources are systematically discussed: overload situations caused by structural system bottle-necks, overload according to the subscriber behaviour and the induction of overload in the telephone network. Section 3.3 deals with the problem of overload detection. Requirements for overload indicators are outlined and examples for typical indicators and their properties are treated. The following two classes of overload indicators are considered: indicators obtained by measurements and actual system state as overload indicator.

In the last part of the chapter a classification of overload control methods is given, where the following classes of control strategies are discussed:

- Enforcement of call completion by means of dynamical adaptation of system parameters
- Reduction of call acceptance
- Global overload control strategies by network management.

CHAPTER 4 MODELLING AND ANALYSIS METHODS

In order to investigate the overload phenomenon and to evaluate the performance of overload control strategies, analytical and simulation methods are employed. They are outlined in this chapter.

Section 4.1 discusses aspects of modelling techniques, especially for overload and overload control performance investigations, while section 4.2 deals with analysis methods, such as time-continuous Markov process, the Kolmogorov equations and the method of imbedded Markov chain, which are used in the later part of the report.

The event-by-event simulation technique for stationary as well as for nonstationary studies is the subject of section 4.3. For the investigation of switching system overload performance, modelling approaches including modelling of nonstationary system loads must be taken into account and a new simulation technique for nonsta-

tionary traffic streams is required. For this purpose, a simulation method for the generalized Poisson process, which is based on a modification of the event-by-event simulation technique, is developed and described in section 4.3.2.

CHAPTER 5 MODELS FOR OVERLOAD CHARACTERIZATION

The characterization of overload traffic and the understanding of the dynamical development of overload situations are required for overload control considerations. Modelling approaches for overload description are dealt with in this chapter.

Section 5.1 contains overload models, where the influence of system reaction on the overload traffic stream is not taken into account. After a brief review of the renewal theory, modelling approaches for the description of call, subprocess and subcall traffic streams in switching systems are presented. Subcall processes are modelled by means of the Generalized Switched Poisson Process (section 5.1.3) for which a renewal approximation is derived.

Considering the feedback effect caused by the customer system interaction, more complex overload models are developed and investigated in section 5.2.

In section 5.2.1 the phenomenon of repeated attempts is modelled (c.f. Fig. 5.12). The model is analyzed by means of a two-dimensional Markov process and a recursive numerical algorithm is developed (Fig.5.13) to compute the state probabilities.

The model in 5.2.2 points out the interdependency between the subcall generation process, the call completion characteristics and the call waiting time according to the customer behaviour. The model is of type $M(t)^{[x]}/M/1$, where the time-dependent call arrival process is described by means of the generalized Poisson process. The batch size distribution which characterizes the subcall process is modelled to be dependent on the system state upon call arrival. Nonstationary system responses to time varying overload traffic patterns which show a hysteresis effect of the dynamical behaviour of a switching system under overload (c.f. Fig.5.21), are numerically obtained.

CHAPTER 6 MODELS FOR OVERLOAD CONTROL STRATEGIES

In this chapter the following classes of overload control strategies are modelled and investigated:

- Control strategies by throttling of call acceptance
- Control strategies by optimizing the system resource usage.

Throttling mechanisms for call acceptance control are developed and analyzed in section 6.1 where two basic call blocking methods are modelled: the blocking scheme according to a two-point control and the gradual, state-dependent call blocking mechanism. In section 6.1.1 the overload control by means of a two-point hysteresis is considered (Figs. 6.1 and 6.2). Based on a state transition diagram (c.f. Fig. 6.3) and a recursion scheme, exact formulae for system characteristics are derived and a dimensioning example for the control strategy is outlined.

Section 6.1.2 deals with the gradual call blocking stragegy. This overload control method is applied to the overload model discussed in section 5.2.2 where the performance degradation of switching systems according to the customer-system interaction is taken into account. The efficiency of the state-dependent call throttling strategy discussed in this chapter is evaluated, whereby the nonstationary system reaction on time-dependent overload traffic patterns with overload control is obtained and analyzed.

Overload control strategies by means of dynamical optimization of system resource usage are modelled in section 6.2. A queueing model for two software machines handling calls in tandem is presented in section 6.2.1, where a dynamical, state-dependent resource allocation scheme is analyzed. Dimensioning aspects are also discussed in order to optimize the call completion.

In section 6.2.2 a compound model for call and subcall handling processes is developed, whereby the dependency of the call completion characteristics on the subcall handling efficiency is modelled in detail (Fig. 6.16). An overload control strategy, the bad-call-interruption scheme (BCI), is presented and investigated. This model is analyzed by means of a two-dimensional Markov process (Fig. 6.18), and the state probabilities are obtained using

a relaxation method. The efficiency of the BCI-control scheme is investigated for stationary cases (Figs. 6.19-21).

CHAPTER 7 CONCLUSION

Concluding remarks are given and major results of the report are summarized.

INHALTSVERZEICHNIS

LITERATURVERZEICHNIS

[1] Feller, W. An Introduction to Probability Theory and Its Applications, Vol. I and II. Wiley & Sons, NewYork, 1968.

[2] Kleinrock, L. Queueing Systems. Vol. I: Theory; Vol. II: Applications. Wiley & Sons, NewYork, 1975/1976.

[3] Gross, D. Harris, C.M. Fundamentals of Queueing Theory. Wiley & Sons, NewYork, 1974.

[4] Takács, L. Introduction to the Theory of Queues. Oxford University Press, NewYork, 1962.

[5] Cox, D.R. Renewal Theory. Methuen & Co., London, 1962.

[6] Cox, D.R. Miller, H.D. The Theory of Stochastic Processes. Methuen & Co., London, 1965.

[7] Kobayashi, H. Modelling and Analysis. An Introduction to System Performance Evaluation. Addison-Wesley Publ.Co., Reading/Mass., 1978.

[8] Cooper, R.B. Introduction to Queueing Theory. Macmillan Co., NewYork, 1972.

[9] Kühn, P. Tabellen für Wartesysteme. Inst. für Nachrichtenvermittlung und Datenverarbeitung, Univ. Stuttgart, 1976.

[10] Kühn, P. Über die Berechnung der Wartezeiten in Vermittlungs- und Rechnersystemen. Dissertationsschrift, Univ. Stuttgart, 1972.

[11] Kühn, P. Analyse zufallsabhängiger Prozesse in Systemen zur Nachrichtenvermittlung und Nachrichtenverarbeitung. Habilitationsschrift, Univ. Stuttgart, 1981.

[12] - Nachrichtenverkehrstheorie - Begriffe. NTG-Empfehlung 1980, Entwurf NTG-0903.

[13] - Programming Languages for Stored Programme Control Exchanges. CCITT Orange Book, Vol. VI.4, Int. Telecomm. Union, Geneva, 1977.

[14] Joel, A.E.Jr. Electronic Switching Central Office Systems of the World. IEEE Press, 1976.

[15] Gerke, P.R. — Rechnergesteuerte Vermittlungssysteme. Springer, Berlin/Heidelberg/NewYork, 1972.

[16] Hills, M.T. — Telecommunications Switching Principles. George Allen & Unwin, London, 1979.

[17] Kühn, P. — Verteilte Mikrorechner-Steuerungen in Nachrichtenvermittlungssystemen - Strukturen, Organisation und Verkehrsanalyse. Sammelband d. Nachr. Koll. der TU Braunschweig: "Nachrichtensysteme - Dienstintegration in künftigen Kommunikationsnetzen". Hrsg. H.L. Hartmann, Teubner, Stuttgart,1982.

[18] Arima, T. et al. — A new Switching Processing Architecture and its Operating System for a Distributed Hierarchy Software Structure. Int. Switching Symp., Paris, 1979, 943-947.

[19] Botsch, D. — Die EWSD-Software - Ein Realzeitprogrammsystem in der höheren Programmiersprache CHILL. Telcom report 2(1979), 184-189.

[20] Suckfüll, H. — Architektur einer neuen Linie digitaler öffentlicher Fernsprechvermittlungen. Telcom report 2(1979), 174-183.

[21] Eberding, H. — Die Software im System EWSD. Telcom report 4(1981), Beiheft Digitalvermittlungssystem EWSD, 13-18.

[22] Carruet, V. Rideau, A. — Software Structure and Methodology. El. Comm, 54(1979) 3, 178-185.

[23] Becker, G. et al. — Call Processing in a Distributed Control System. Proc. Int. Conf. on Comm., Seattle,1980,pp.46.4.

[24] Cox, I.E. et al. — A Digital Switch for Wide Range of Application. Proc. Int. Conf. on Comm., Seattle,1980,pp.46.1.

[25] - — System 12 - Das Digital-Vermittlungssystem. Technische Beschreibung. Standard Elektrik Lorenz, 1981.

[26] Katzschner, L. Weisschuh, H. — Das Software-Strukturkonzept eines rechnergesteuerten Vermittlungssystems. NTG/GI Tagung: "Struktur und Betrieb von Rechensystemen", Ulm, 1982; NTG Fachberichte, Band 80.

[27] Carestia, P.D. Hudson, F.S. — No. 4 ESS: Evolution of the Software Structure. Bell Syst. Tech. J. 60(1981) 6, 1167-1201.

[28] Bruce, R. A. et al. — No. 4 ESS - Evolution of a Digital Switching System. IEEE Trans. Comm. 27 (1979)7, 1001-11.

[29] Andrews,F.T.,Jr. Smith,Wm.Br. — No. 5 ESS - Overview. Int. Switching Symp., Montreal, 1981.

[30] Davis, J. H. et al. — No. 5 ESS - System Architecture. Int. Switching Symp., Montreal, 1981.

[31] Bauman, S. M. et al. — No. 5 ESS - Software Design. Int. Switching Symp., Montreal, 1981.

[32] Romoeuf, L. — Modelling a Stored Program Controlled Telephone Switching System: Evaluating a Regulation Method for Traffic Overloads. Symp. on Meas., Mod. and Eval. of Comp. Systems, North Holl. Publ. Co., 1977.

[33] Forys, L. J. — Modelling of SPC Switching Systems. Proc. of the 1st. ITC-Seminar on Modelling of SPC Exchanges and Data Networks, Delft, Oct. 1977, 83-100.

[34] Wizgall, M. — Über Architektur, Betriebsweise und Verkehrsverhalten der Steuerung einer rechnergesteuerten Vermittlungsstelle. Dissertationsschrift, Universität Stuttgart, 1980.

[35] Dietrich, G. — Verkehrsmodelle für zentralgesteuerte Vermittlungssysteme. El. Nachrichtenwesen 50 (1975)1, 30-36.

[36] Dietrich, G. Salade, R. — Teilbelegungstreue Simulation der Steuerung von Vermittlungssystemen. El. Nachrichtenwesen 52 (1977)1, 61-68.

[37] Karlander, B. — Control of Central Processor Load in a SPC System. Proc. 7th I.T.C., Stockholm, 1973.

[38] Schoute, F. C. — Optimal Control and Call Acceptance in a SPC Exchange. Proc. 9th I.T.C., Torremolinos, 1979.

[39] Briccoli, A. — Comparison of Regulation Methods for Traffic Overloads in SPC Systems. Proc. 9th I.T.C., Torremolinos, 1979.

[40] Somosa, M. A. Guerrero, A. — Dynamic Processor Overload Control and its Implementation in Certain Single Processor and Multiprocessor SPC Systems. Proc. 9th I.T.C., Torremolinos, 1979.

[41] Miller, S. E. Service Continuity Planning. Proc. Nat. Telecomm. Conf., New Orleans, Nov/Dec. 1981, pp. B4.1.

[42] Hashida, D. et al. Congestion Mechanisms and Overload Control in Telephone Networks. Proc. 9th I.T.C., Torremolinos, 1979.

[43] Arthurs, E. Stuck, B. W. Controlling Overload in a Digital System. SIAM J. Alg. Disc. Meth., 1 (1980), 2.

[44] Borchering,J.W. et al. Coping with Overloads. Bell Lab. Record, 7/8 (1981), 183-185.

[45] Hentschke, S. Predictive Processor Overload Control Strategies for SPC Switching Systems. NTZ-Archiv, 3 (1981), 5, 121-127.

[46] Harvey, C. Griffiths,P.H. Controlling Overloads in Multiprocessor Systems. Proc. 4th Int. Conf. on Soft. Eng. for Telecomm. Switching Systems, 1981, 167-171.

[47] Tran-Gia, P. Modelling and Analysis of Software Resources in Modular SPC-Switching Systems - Some Aspects of Dimensioning and Overload Control. Proc. 4th Int. Conf. on Soft. Eng. for Telecomm. Switching Systems, 1981, 172-176.

[48] Radant, E. Simulation von Steueraufrufprozessen zur Leistungsanalyse rechnergesteuerter Fernsprechvermittlungssysteme. Diplomarbeit Nr. 24, FG Nachrichtentechnik, Universität Siegen, 1981.

[49] Kuczura, A. Queues with Mixed Renewal and Poisson Inputs. Bell Syst. Tech. J. 51 (1972)6, 1305-1326.

[50] Kuczura, A. The Interrupted Poisson Process as an Overflow Process. Bell Syst. Tech. J. 52 (1973)3, 437-448.

[51] Heffes, H. On the Output of a GI/M/N Queueing System with Interrupted Poisson Input. Operations Research 24 (1976)3, 530-542.

[52] Heffes, H. A Class of Data Traffic Processes - Covariance Function Characterization and Related Queueing Results. Bell Syst. Tech. J. 59 (1980)6, 897-929.

[53] Yechiali, U. Naor, P. Queueing Problems with Heterogeneous Arrivals and Service. Operations Research 19 (1971), 722-734.

[54] Fond, S. Ross, S. M. A Heterogeneous Arrival and Service Queueing Loss Model. Naval Res. Log. Quart. 25 (1978), 483-488.

[55] Myskja, A. A Study of Telephone Traffic Arrival Processes with Focus on non-stationary Cases. University of Trondheim, Norwegian Institute of Technology, 1981.

[56] Jüchter, H. Simulation instationärer Vorgänge in Warteschlangensystemen. Diplomarbeit Nr. 25, FG Nachrichtentechnik, Universität Siegen, 1981.

[57] Bieker, B. Analyse des Steueraufrufverkehrs in rechnergesteuerten Vermittlungssystemen mit dem geschalteten Poisson-Prozeß. Studienarbeit Nr. 27, FG Nachrichtentechnik, Universität Siegen, 1982.

[58] Nesenbergs, M. A Hybrid of Erlang B and C Formulas and its Applications. IEEE Trans. Comm., 27 (1979)1, 59-68.

[59] Jonin, G. L. Sedol, J. J. Telephone Systems with Repeated Calls. Proc. 6th I.T.C., Munich, 1970.

[60] Schneps-Schneppe, M. The Effect of Repeated Calls on Communication System. Proc. 6th I.T.C., Munich, 1970.

[61] Bretschneider, G. Repeated Calls with Limited Repetition Probability. Proc. 6th I.T.C., Munich, 1970.

[62] Gosztony, G. Comparison of Calculated and Simulated Results for Trunk Groups with Repeated Attempts. Proc. 8th I.T.C., Melbourne, 1976.

[63] Grillo, D. Telephone Network Behaviour in Repeated Attempts Environment - A Simulation Analysis. Proc. 9th I.T.C., Torremolinos, 1979.

[64] Macfadyen,N.W. Statistical Observation of Repeated Attempts in the Arrival Process. Proc. 9th I.T.C., Torremolinos, 1979.

[65] Myskja, A. Aagesen, F. A. On the Interaction between Subscribers and a Telephone System. Proc. 8th I.T.C., Melbourne, 1976.

[66] Liu, K. S. Direct Distance Dialing: Call Completion and Customer Retrial Behaviour. Bell Syst. Tech. J., 59 (1980)3, 295-311.

[67] Posner, M. Single Server Queues with Service Time dependent on Waiting Time. Operations Research, 21 (1973), 610-616.

[68] Rosenshine, M. Queues with State-dependent Service Times. Transpn. Res., 1(1967), 97-104, (Pergamon Press).

[69] Harris, C. M. Some Results for Bulk-arrival Queues with State-dependent Service Times. Management Science, 5 (1970), 313-326.

[70] Tran-Gia, P. Van Hoorn, M.H. Dependency of Service Time on Waiting Time in Switching Systems - A Queueing Analysis with Aspects of Overload Control. Universität Siegen - Freie Universität Amsterdam, 1982.

[71] Manfield, D.R. Tran-Gia, P. Queueing Analysis of Scheduled Communications Phases in Distributed Processing Systems. Proc. 8th Symp. on Comp. Perf. Modelling, Meas. and Eval., Amsterdam, Nov. 1981, 233-250.

[72] Stüken, R. Zur Optimierung der Wartezeiten in rechnergesteuerten Vermittlungssystemen mit adaptiven Scheduling-Verfahren. Diplomarbeit Nr. 28, FG Nachrichtentechnik, Universität Siegen, 1982.

[73] Little, J.D.C. A Proof of the Queueing Formula $L=\lambda W$. Operations Research, 9 (1961), 383-387.

[74] Herzog, U. Woo, L. Chandy, K.M. Solution of Queueing Problems by a Recursive Technique. IBM J. Res. Dev. 19(1975)3, 295-300.

[75] Törnig, W. Numerische Mathematik für Ingenieure und Physiker, Band 1&2. Springer, Berlin/Heidelberg/New York, 1979.

Verzeichnis der wichtigsten Abkürzungen und Formelzeichen

λ	Ankunftsrate Die Indizierung erfolgt modellspezifisch und nach Art der Anforderung (Ruf, Teilprozeß oder Teilruf)
ε,μ	Enderate - von Bedienungseinheiten -
BCI	bad-call-interruption : Überlastabwehrstrategie mit Zwangsauslösung nichtkomplettierbarer Rufe
CCB	call-control-block : rufbezogener Datenblock
LT	Laplace-Transformation
LST	Laplace-Stieltjes-Transformation
P{...}	Wahrscheinlichkeit für {...}
SM	Software-Maschine
SPP	switched Poisson-process : geschalteter Poisson-Prozeß
ÜWD	Übergangswahrscheinlichkeitsdichte
ZV	Zufallsvariable

Andere modellspezifische Abkürzungen und Formelzeichen werden in den entsprechenden Unterkapiteln aufgeführt.

Notation für Zufallsvariablen und Transformationsmethoden :

T	Zufallsvariable (ZV) - hier z.B. zeitbezogen -
$E[T^n]$	n-tes gewöhnliches Moment der ZV T $E[T]$: Mittelwert der ZV T
T^v	Vorwärts-Rekurrenzzeit der ZV T
T^r	Rückwärts-Rekurrenzzeit der ZV T
$F(t) =$	$P\{T \leq t\}$ Verteilungsfunktion der Zufallsvariable T
$f(t) =$	$dF(t)/dt$ Verteilungsdichtefunktion der Zufallsvariable T
$\Phi(s) =$	$LT\{f(t)\} = LST\{F(t)\}$ Laplace-Transformierte von f(t) bzw. Laplace-Stieltjes-Transformierte von F(t).

1. EINLEITUNG

1.1 Zur Überlast- und Überlastabwehrproblematik

Die Problematik der Systemüberlastung und Überlastabwehr begleitet die Entwicklung aller Generationen von Fernsprechvermittlungssystemen. Bei der konventionellen Vermittlungstechnik war die Dringlichkeit der Entwicklung, Erprobung und Implementierung von Überlastabwehrmaßnahmen noch nicht gegeben, da die herkömmlichen Vermittlungssysteme häufig eine für Überlast relativ unempfindliche dezentrale Steuerungsstruktur aufweisen.

Infolge des technologischen Fortschritts, der zur Entwicklung rechnergesteuerter Vermittlungssysteme führte, wird die Überlaststeuerung zu einem der zentralen Probleme, welche die Leistungsfähigkeit eines Vermittlungssystems entscheidend beeinflussen. Durch die Zunahme der Anzahl von Fernsprechanschlüssen und durch die Einführung neuer Dienste und Leistungsmerkmale im Fernsprechwesen (Tastwahl, automatische Anrufwiederholung, Anrufumleitung, Kurzwahl, ...) wird die Überlastproblematik noch komplexer. Die Entwicklung neuartiger Systemstrukturen - z.B. die hard- und softwaremäßige Aufteilung vermittlungstechnischer Funktionen in autonome Module, die teil- und dezentrale Steuerungsstruktur usw.- trägt ebenfalls dazu bei, daß die Gewährleistung der Funktionsfähigkeit rechnergesteuerter Vermittlungssysteme in Überlastsituationen schwieriger wird.

Aspekte des Überlastproblems müssen in der Planungs-, Entwicklungs- und Betriebsphase eines Vermittlungssystems berücksichtigt werden. Da das Verkehrsgeschehen Zufallsprozesse darstellt, werden für Untersuchungen zur dynamischen Überlastentwicklung und zur Leistungsfähigkeit von Überlastabwehrmaßnahmen Methoden der stochastischen Systemanalyse herangezogen.

In der Literatur findet man eine Anzahl von Untersuchungen, die sich mit der Problematik der Überlaststeuerung befassen. Ein großer Teil dieser Arbeiten präsentiert Implementierungen von

Überlastabwehrstrategien in realen Systemen, wobei Simulationsstudien und Messungen diskutiert werden [34, 35, 36]. Andere Untersuchungen befassen sich mit Aspekten zur verkehrstheoretischen Modellbildung und Analyse der Überlast- und Überlastabwehr-Problematik [32, 33, 46]. Einige Grundmodelle werden in diesen Studien, in denen systemspezifische Überlastabwehrmethoden untersucht werden [37-40, 43-45], behandelt.

1.2 Übersicht über die Arbeit

In der vorliegenden Arbeit werden im Zusammenhang mit unterschiedlichen System- und Steuerungsstrukturen rechnergesteuerter Fernsprechvermittlungssysteme folgende Teilaspekte diskutiert:

- Die Erkennung von Überlastsituationen: Überlastindikatoren, zeitliche Anforderung der Überlasterkennung.
- Überlastabwehrstrategien: dynamische Wirkungsweise, Wirkbreite, Einfluß der Systemparameter, Leistungsfähigkeit der Abwehrmethoden.

Die Untersuchungen basieren auf verkehrstheoretischen Modellen, wobei aus einigen Grundmodellen komplexere Modelle gewonnen werden, welche

- Verkehrsströme in rechnergesteuerten Fernsprechvermittlungssystemen beschreiben.
- die Entwicklung der Überlastsituationen darstellen.
- qualitative und quantitative Leistungsbeurteilung entwickelter Überlaststeuerungsstrategien ermöglichen.

In Kap. 2 werden Grundaufgaben und Steuerungsstrukturen rechnergesteuerter Fernsprechvermittlungssysteme erläutert. Die Probleme der Interprozeß- und Interprozessor-Kommunikation sowie der Beschreibung von Verkehrsströmen in Vermittlungssystemen, insbesondere in modular strukturierten Systemen, werden diskutiert.

Kap. 3 befaßt sich mit der Problematik der Überlast und Überlast-

abwehr in Vermittlungssystemen. Nach allgemeinen Definitionen werden Ursachen und Entwicklungen von Überlastsituationen behandelt. Abschließend wird der Versuch unternommen, die weitläufigen Aspekte der Überlastindikatoren und Überlastabwehrmaßnahmen in einer klassifizierenden Darstellung zu ordnen.

Das 4. Kapitel gibt einen Überblick über die in der vorliegenden Arbeit angewendeten Methoden zur verkehrstheoretischen Modellbildung und -analyse. Es handelt sich hierbei um analytische und simulative Untersuchungsverfahren, die in den Kap. 5 und 6 Anwendung finden. Neben den bekannten Methoden für die Analyse kontinuierlicher und diskreter Markoff-Prozesse sowie der häufig benutzten Methode der zeittreuen Simulation stationärer Zufallsprozesse wird eine neue Methode, die sog. Vorwärts-Integral-Methode, vorgestellt, welche zur Simulation instationärer Ankunftsprozesse - insbesondere des verallgemeinerten Poisson-Prozesses - entwickelt wurde.

Im Anschluß an die einführenden Kapitel 1-4 werden in Kap. 5 und 6 Verkehrsmodelle entwickelt und analysiert, mit denen die Eigenschaften von Überlastverkehrsströmen charakterisiert werden können sowie eine quantitative Leistungsbeurteilung von Überlastabwehrmaßnahmen ermöglicht wird.

In Kap. 5 werden Verkehrsmodelle und Modellkomponenten für die Beschreibung des Überlastverkehrs sowie der Überlastsituationen behandelt. Eine Erneuerungsapproximation für den geschalteten Poisson-Prozeß, der zur Beschreibung von Ankunftsprozessen mit starken Schwankungen Anwendung findet, wird vorgestellt. Weitere Modellansätze hinsichtlich der Teilnehmer-System-Interaktion werden untersucht, welche insbesondere rückwirkungsbehaftet sind, d.h. den Einfluß des Vermittlungssystems auf die Eingangsprozesse berücksichtigen. Es handelt sich hierbei um die Modellbildung des Rufwiederholungseffektes und der wartezeitabhängigen Rufkompletierungscharakteristik eines Vermittlungssystems.

Überlastabwehrmethoden werden in Kap. 6 anhand einiger Grund-

modelle untersucht. Neben den Grundmechanismen zur Drosselung der Rufannahme - der Zweipunkt-Regelung und der graduellen Blockierung - werden verschiedene Überlastabwehrmaßnahmen zur optimalen Ausnutzung der Systemkapazität vorgestellt. Mit den gewonnenen Ergebnissen aus den Verkehrsuntersuchungen wird die Leistungsfähigkeit dieser Abwehrmethoden im stationären und instationären Falle diskutiert.

2. GRUNDAUFGABEN UND STEUERUNGSPRINZIPIEN RECHNERGESTEUERTER FERNSPRECHVERMITTLUNGSSYSTEME

In diesem Kapitel werden Struktur und Betriebsmerkmale rechnergesteuerter Vermittlungssysteme sowie die Beschreibung vermittlungstechnischer Prozesse erläutert und einige Aspekte diskutiert, welche zum Verständnis der im nächsten Kapitel behandelten Überlastproblematik beitragen sollen.

2.1 Grundaufgaben eines Vermittlungssystems

Ein Fernsprechvermittlungssystem hat die Aufgabe, eine zeitweilige Verbindung zwischen zwei Teilnehmern (Tln.) herzustellen, wobei der eine Tln. - der rufende Tln. A - dem Vermittlungssystem die Zielinformation des anderen Tln. - der gerufene Tln. B - mitteilt. Dabei können die an der Verbindung beteiligten Teilnehmer an dieselbe Vermittlungsstelle oder an verschiedene Vermittlungsstellen eines Fernsprechnetzes angeschlossen sein.

Das Vermittlungssystem stellt Vermittlungseinrichtungen zur Verfügung, mit denen Steuerinformationen (in Form von Wählziffern oder internen vermittlungstechnischen Signalen) verarbeitet und die Nutzinformation (das Gespräch zwischen Tln. A und B) übertragen werden. Im folgenden werden die an einer Verbindung beteiligten Vermittlungsfunktionen und -prozesse näher erläutert.

2.1.1 Vermittlungsfunktionen

Eine erfolgreiche Verbindung besteht prinzipiell aus drei Phasen: der Verbindungsaufbauphase, der Gesprächsphase und der Auslösungsphase. Die Steuerung sowie die Vermittlungseinrichtungen werden während der Verbindungsaufbauphase am intensivsten beansprucht. Am Beispiel des Aufbaus einer internen Verbindung - d.h. beide Teilnehmer sind an dieselbe Vermittlungsstelle angeschlossen - werden einige Grundaufgaben eines Vermittlungssystems dargestellt (s. Bild 2.1):

- Überwachung des Teilnehmerzustandes:
  Das Vermittlungssystem überwacht ständig den aktuellen Zustand (z.B. abgehoben oder aufgelegt) aller angeschlossenen Teilnehmer, Verbindungsleitungen und Hilfseinrichtungen, z.B. durch eine taktgesteuerte Abtastprozedur.
- Steuerung des Wahlvorganges:
  Liegt ein Verbindungswunsch vor, so wird der Tln. A identifiziert und auf seine Berechtigung hin überprüft. Ein Tongenerator wird zur Wahlaufforderung bereitgestellt. Bild 2.1 zeigt

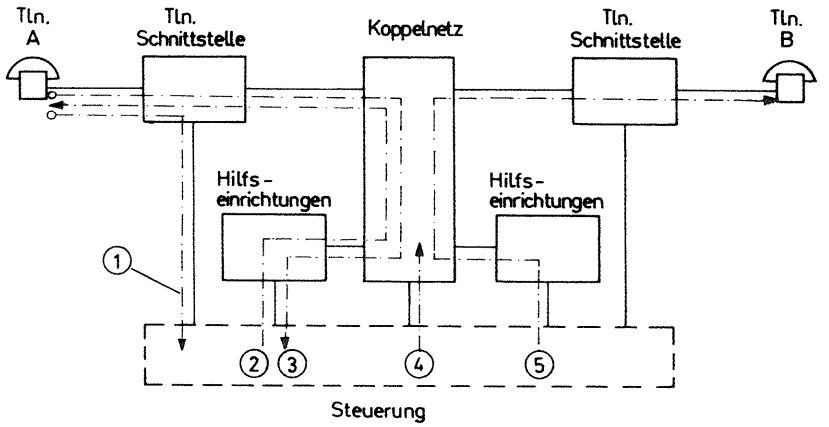


Bild 2.1 *Steuervorgänge beim Aufbau einer internen Verbindung.*

1. *Feststellung des Verbindungswunsches und Identifizierung des A-Tln.*
2. *Wahlaufforderung*
3. *Empfang und Auswertung der Wählinformation*
4. *Markierung ; Einstellen des Koppelnetzes für Verbindung A - B*
5. *Anschalten des Ruftons zum B-Tln.*

ein Realisierungsbeispiel dieser Aufgabe, indem ein Weg durch das Koppelnetz gesucht wird, der eine Verbindung des A-Tln. mit dem Tongenerator zuläßt. In der Darstellung nach Bild 2.1 befindet sich der Tongenerator im Modul der vermittlungstechnischen Hilfseinrichtungen. Parallel dazu wird ein Ziffernempfänger zur Verfügung gestellt. Beim Ziffernempfang werden die Zeitdauer zwischen den einzelnen Ziffern überwacht, die Wählinformation auf Vollständigkeit hin überprüft und das Wahlende registriert.

- Durchschaltung der Verbindung:
  Befindet sich der B-Tln. im Freizustand, so wird ein Weg zwischen Tln.A und Tln.B gesucht. Bei erfolgreicher Wegesuche wird die Verbindung markiert, d.h. der Weg wird für die aufzubauende Verbindung reserviert. In rechnergesteuerten Vermittlungssystemen geschieht dies softwaremäßig im Datenbereich für die Koppelnetz-Zustandsspeicherung. Parallel dazu werden der Rufton zum Tln.B und der Freiton zum Tln.A durchgeschaltet. Meldet sich der Tln.B. so werden alle Töne abgeschaltet und der Sprechweg durchgeschaltet. Die Verbindung tritt dann in die Gesprächsphase ein.

Neben den hier beschriebenen Aufgaben werden in der Gesprächs- und der Auslösungsphase weitere vermittlungstechnische Steuerungsaufgaben im Echtzeitbetrieb durchgeführt, die hier nicht näher betrachtet werden.

In der Regel steuert ein Vermittlungssystem gleichzeitig mehrere Verbindungen, die sich in unterschiedlichen Phasen befinden. Die Verarbeitung der von den aktiven Verbindungen erzeugten Steuerungsaufrufe unterliegt strengen Echtzeit-Anforderungen. Man findet deshalb in Vermittlungssystemen nahezu sämtliche Betriebsmerkmale von Echtzeit-Rechnersystemen zur Prozeßautomatisierung, z.B. Mehrprogrammbetrieb, Mehrrechnerbetrieb, Funktions- und/oder Lastteilungsprinzip usw. mit den begleitenden Problemen der Prozeßsynchronisation, der schnellen Ein/Ausgabemechanismen, usw. vor.

In einem rechnergesteuerten Vermittlungssystem kann das oben beschriebene Szenario eines Verbindungsaufbaus, abhängig von der Systemarchitektur (s.Kap. 2.2), von einem oder mehreren Rechnern gesteuert werden. Man kann jedoch bei allen Steuerungsstrukturen folgende Grundfunktionen erkennen:

- Steuerung peripherer Schnittstellen:
  Die Peripherie einer Vermittlungsstelle besteht aus Teilnehmer-Apparaten, Verbindungsleitungen, Hilfseinrichtungen (Ton-Generatoren, interne Zeitüberwachungseirichtungen,...), dem Koppelnetz sowie verwaltungstechnischen Einrichtungen. Der augenblickliche Zustand einzelner Einrichtungen wird durch Abtastvorgänge (scanning) überwacht. Die hier vorliegenden Zustandsinformationen oder Zustandsänderungen sind noch in einer physikalischen Form dargestellt, z.B. Spannungen oder Ströme (Tln.-Zustand, Impulsdarstellung für Wählziffern,...). Diese Signale werden für die nächste Steuerungsstufe, die Signalisierung, zu gültigen telephonischen Ereignissen (Belegungsversuche, zusammenhängende Ziffernfolge für die Wegesuche usw.) verarbeitet.

- Signalisierung:
  Es wird unterschieden zwischen Teilnehmer- und Amtssignalisierung, ferner zwischen ankommender und abgehender Signalisierung. Da an einer Verbindung im Fernsprechnetz mehrere Vermittlungsstellen beteiligt sein können, müssen vermittlungstechnische Steuersignale zwischen den Teilnehmern und dem Vermittlungssystem sowie zwischen Vermittlungssystemen ausgetauscht werden. Dies geschieht entlang einer Kette von ankommenden und abgehenden Signalisierungen. Die Signalisierung stellt die Realisierung eines Protokolls zwischen Vermittlungseinrichtungen bzw. -prozessen dar.

- Rufverarbeitung:
  Diese zentrale Aufgabe eines Vermittlungssystems beinhaltet alle Funktionen, die für den Aufbau, die Überwachung und die Auflösung einer Verbindung erforderlich sind, z.B. Ziffern-

auswertung, Wegesuche, Durchschaltung des Sprechweges. Die Rufverarbeitung umfaßt ebenfalls die Realisierung unterschiedlicher Leistungsmerkmale, z.B. Kurzwahl, Anrufumleitung usw., die in modernen rechnergesteuerten Vermittlungssystemen implementiert sind.

- Systemüberwachung und Wartung:
  Hierzu gehören die Testroutinen und Diagnoseprüfungen, die Analyse der Fehlermeldungen sowie die Rekonfiguration und Wiederherstellung von Systemfunktionen nach der Fehlerbehandlung. Die Realisierung dieser Funktion geschieht, abhängig von der Dringlichkeit und von der Häufigkeit der jeweiligen Testroutine, sowohl in Hard- als auch in Software. Die Systemüberwachung spielt für die Überlastabwehr eine entscheidende Rolle, da die Gültigkeit und die Aktualität der Überlastindikatoren (vgl.Kap.6) davon abhängig sind, wie leistungsfähig die Fehlerdiagnose und die Routine-Meßvorgänge sind.
- Verwaltung:
  Beispiele dafür sind das Beschalten und das Freischalten von Teilnehmeranschlüssen sowie die Zuordnung zwischen Rufnummer und Anschlußposition von Teilnehmern. Ferner beinhaltet diese Aufgabe die Verwaltung von Verkehrslenkungstabellen und die Gebührenerfassung. Die Verwaltungsaufgaben sind z.T. aufschiebbare Arbeiten, d.h. falls Überlast bezüglich dringlicheren vermittlungstechnischen Hauptfunktionen auftritt, werden Verwaltungsaufgaben mit einer niedrigeren Priorität ausgeführt.

2.1.2 Vermittlungsprozesse

Eine Verbindung, zu welcher der Aufbau, die Gesprächsphase und deren Auflösung gehört, stellt für das Vermittlungssystem eine Folge von Signalen bzw. Ereignissen dar (Belegungswunsch, Wählziffern,...). Auf ein Ereignis muß das System, abhängig vom jeweiligen Zustand der Verbindung, mit einer festgelegten Reihe von Aktionen und Signalen zur Steuerung der Verbindung reagieren. Da die Menge aller Zustände einer Verbindung endlich ist, kann eine Verbindung (in den späteren Kapiteln als "Ruf" bezeichnet) mit einem Vermittlungsprozeß, der einem endlichen Automaten (FSM: finite state machine) entspricht, vollständig charakterisiert werden.

Da eine Verbindung von mehreren vermittlungstechnischen Einrichtungen gesteuert wird und unterschiedliche Vermittlungsfunktionen benötigt, ist es zweckmäßig, den gesamten Vermittlungsprozeß in mehrere funktionsorientierte Teilprozesse aufzuteilen. Ein Teilprozeß enthält im Vergleich zum gesamten Vermittlungsprozeß weniger Zustände und Ereignisse. Dadurch sind Teilprozesse überschaubarer und softwaremäßig leichter implementier- und dokumentierbar.

Da mehrere verbindungsbezogene Teilprozesse an der Steuerung einer Verbindung beteiligt sein können und die Teilprozesse häufig in verschiedenen Steuerungseinheiten lokalisiert sind, ist die Leistungsfähigkeit der Interprozeß- und Interprozessor-Kommunikation ein entscheidender Faktor für die Funktionsfähigkeit des Gesamtsystems.

Bild 2.2 zeigt die wesentlichen vermittlungstechnischen Teilprozesse, die zur Steuerung einer Verbindung aktiviert werden müssen. Die Teilprozesse werden hier für abgehende und ankommende Verbindungen (bzw. entspringende und endende Verbindungen) aufgeteilt. In einigen Realisierungen werden die abgehenden und ankommenden Verbindungssteuerungsprozesse zusammen in einem gesamten Prozeß implementiert, der in der Lage sein muß, mit unterschiedlichen Signalisierungssystemen zu kommunizieren. Die physikalischen Signale aus der Peripherie (z.B. Teilnehmer-Apparat) werden vom Teil-

nehmerschnittstellenprozeß auf ihre Gültigkeit hin überprüft und als physikalische Ereignisse interpretiert, die an der physikalischen Schnittstelle mit dem Signalisierungsprozeß vorliegen. Dieser Prozeß verarbeitet die physikalischen Ereignisse (z.B. Schleifenzustandsänderung beim Teilnehmer durch Abheben oder Auflegen des Hörers, einzelne Wählziffern,...) und formt daraus gültige telephonische Ereignisse (z.B. Belegen, gültige Ziffernfolge,...), welche die Ereignisse für den Verbindungssteuerungsprozeß darstellen. Die hier beschriebene Verarbeitung der Ereignisse erfolgt in beiden Richtungen (s. Bild 2.2).

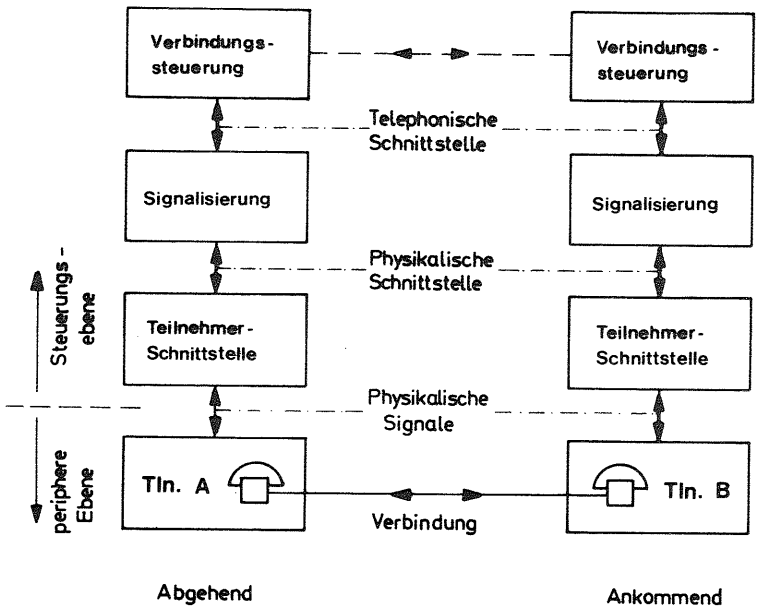


Bild 2.2 Vermittlungsprozesse.

2.2 Steuerungsprinzipien und Architektur

Bei den unterschiedlichen Konfigurationen und Steuerungsprinzipien von rechnergesteuerten Fernsprechvermittlungssystemen ist es nicht leicht, eine Klassifizierung der im Betrieb oder in der Entwicklung befindlichen Systeme hinsichtlich der Hardware- und der Software-Strukturen sowie der Steuerung vorzunehmen [14-17]. In [14] wird eine standardisierte Form zur Strukturbeschreibung von Vermittlungssystemen vorgestellt.

Einige strukturelle Merkmale sowie Organisationsprinzipien der Steuerung können zur Klassifizierung herangezogen werden:

Steuerung	· konzentriert · verteilt
Hardware-Struktur:	· zentral · dezentral · gemischt zentral/dezentral
Software-Struktur:	· nicht-modular · modular

Man findet in einem System mit zentraler Struktur häufig das konzentrierte Steuerungsprinzip, während das in modernen Fernsprechvermittlungssystemen [18-31] oft angewendete Prinzip der verteilten Steuerung eine dezentrale bzw. gemischt zentral/dezentrale Systemstruktur voraussetzt. Im folgenden werden einige typische Systemarchitekturen vorgestellt.

2.2.1 Zentrale Struktur mit konzentrierter Steuerung

Bei dieser Steuerung konzentrieren sich alle vermittlungstechnischen Funktionen in einer Steuerungseinheit, die aus einem oder mehreren gleichartigen Rechnern besteht (Bild 2.3). In dieser zentralen Steuerungseinheit befinden sich alle Programme, die für die Verbindungssteuerung benötigt werden, sowie die Zustandsspei-

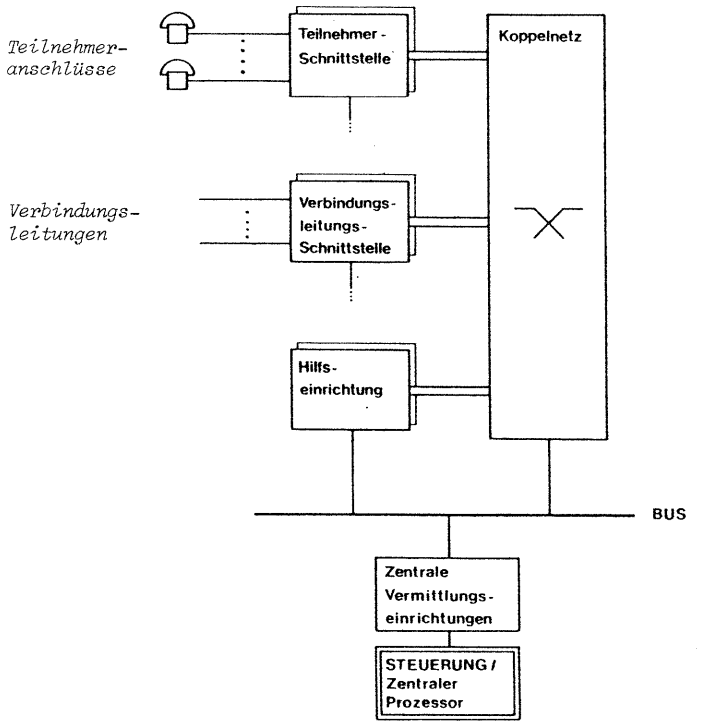


Bild 2.3 *Zentrale Systemstruktur mit konzentrierter Steuerung.*

cherung des gesamten Systems, d.h. ein komplettes Abbild der aktuellen Zustände von Teilnehmern, Verbindungsleitungen, dem Koppelnetz und der systemeigenen peripheren Einrichtungen.

Diese Steuerungsart erfordert eine große Speicher- und Rechenkapazität sowie eine hohe Zuverlässigkeit des Steuerrechners, welcher aus Sicherheitsgründen in der Regel gedoppelt wird. Die Leistungsfähigkeit des Systems wird durch die Kapazität der Steuereinheit

bestimmt, die für die Überlast-Betrachtung den Engpaß des Systems bildet. Die zentrale Steuereinheit muß in der Regel für den Endausbau des Vermittlungssystems dimensioniert werden. Die zentrale Systemstruktur verursacht, bedingt durch die Konzentration der gesamten Intelligenz des Systems, einen intensiven Steuerdatentransport zwischen der peripheren Ebene und der zentralen Steuereinheit. Diese Aufgabe erfordert einen leistungsfähigen Ein/Ausgabemechanismus für vermittlungstechnische Signale und Befehle.

Bild 2.3 zeigt eine übliche Realisierung des Steuerdatentransfers, indem alle Steuersignale durch ein Bussystem von der Peripherie zur Steuerung und umgekehrt transportiert werden.

2.2.2 Dezentrale Struktur mit verteilter Steuerung

Die technologischen Fortschritte, insbesondere die Entwicklung von Mikrorechnern und schnellen Speichern in der Hardware und die Verfügbarkeit von Entwicklungs- und Beschreibungsmethoden in der Software, ermöglichen eine wirtschaftliche Verlagerung der Intelligenz in die periphere Systemebene. Die Systemfunktionen werden auf mehrere Steuereinheiten verteilt, wobei jede Steuereinheit (Rechner) ausschließlich für ein Modul zuständig ist, z.B. Teilnehmer- oder Verbindungsleitungsschnittstelle usw. (Bild 2.4). Die Module können wegen ihrer autonomen Steuerung beim Ausbau der Vermittlungsstelle nach Bedarf hinzugefügt werden. Dadurch werden die Realisierung und die Anpassung verschiedener neuer Dienste oder Signalisierungssysteme erleichtert. Bei der Erweiterung eines Teilsystems wird der Normalbetrieb des vorhandenen Systems nicht wesentlich beeinflußt. Überdies bietet die verteilte Steuerung eine erhöhte Ausfallsicherheit und eine verringerte Wirkbreite von Störungen.

Man findet bei der dezentralen Systemstruktur [18-26] häufig eine Mischung zwischen Funktions- und Lastteilungsprinzipien. Die Vermittlungsfunktionen werden je nach Aufgabe auf die dezentralen Module verteilt (Funktionsteilung). Einige wichtige Vermittlungsfunktionen (z.B. Verbindungssteuerung) werden von mehreren

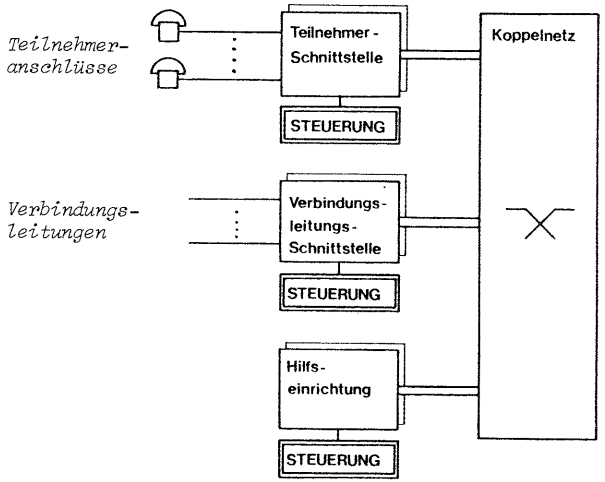


Bild 2.4 *Dezentrale Systemstruktur mit verteilter Steuerung.*

Steuereinheiten übernommen, die nach dem Lastteilungsprinzip operieren.

Da die dezentralen Steuereinheiten autonome Datenbereiche besitzen und keine Steuereinheit das gesamte Abbild der Systemperipherie in seinem Speicherbereich enthält, muß eine funktionsfähige Interprozessorkommunikation für den Datenaustausch gewährleistet werden.

Das für die Betrachtung von Überlastsituationen und Überlastabwehr wichtige Problem der Interprozeß- und Interprozessorkommunikation, das in völlig dezentraler Systemstruktur eine besonders kritische Rolle spielt, wird in Kap. 2.3.2 näher untersucht.

2.2.3 Gemischt zentral/dezentrale Struktur mit verteilter Steuerung

Diese Struktur findet Anwendung in den meisten rechnergesteuerten Vermittlungssystemen. Ein Teil der Systemintelligenz wird hier in die periphere Ebene verlagert. Es handelt sich hierbei um vermittlungstechnische Funktionen, die ohne großen Kommunikationsaufwand

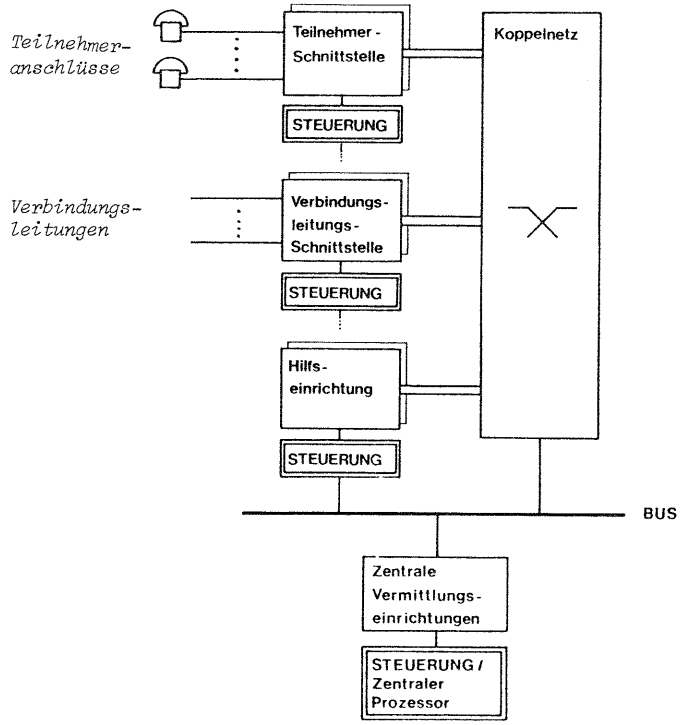


Bild 2.5 *Gemischt zentral/dezentrale Systemstruktur mit verteilter Steuerung.*

mit der zentralen Steuereinheit extern verarbeitet werden können. Beispiele hierfür sind das Abtasten von Teilnehmern und von Verbindungsleitungen, der Ziffernempfang und die Überwachung des Wahlvorganges, die Ablaufsteuerung der Ein/Ausgabe von Steuersignalen usw. (s. Bild 2.5).

Eine Variante dieser Systemarchitektur [20, 23, 25, 26] ist eine Systemkonfiguration mit dezentraler Hardware-Struktur, bei der eine hierarchische Steuerungsstruktur realisiert ist. Während einige Steuereinheiten bestimmten Teilnehmergruppen, Verbindungsleitungen oder Hilfseinrichtungen fest zugeordnet sind, übernehmen andere Steuereinheiten bzw. Rechner einer höheren Steuerebene die zentralen Vermittlungsfunktionen (z.B. Signalisierung, Verbindungssteuerung, ...). Abhängig von der Intensität des Steuerdatenverkehrs wird die Kommunikation zwischen den Steuereinheiten organisiert. Sie kann entweder über das Koppelnetz mittels fest zugeordneter bzw. vermittelter Kanäle erfolgen oder, wie in Bild 2.5 dargestellt, mit einem Bussystem realisiert werden.

2.3 Interprozessor- und Interprozeß-Kommunikation

Aufgrund der Aufteilung von Vermittlungsfunktionen in autonome Module, die sowohl in der Soft- als auch in der Hardware moderner Vermittlungssysteme realisiert wird, erhöht sich der Steuerungsaufwand für den Austausch von vermittlungstechnischen Signalen zwischen Steuereinheiten und zwischen autonomen Software-Modulen innerhalb einer Steuereinheit.

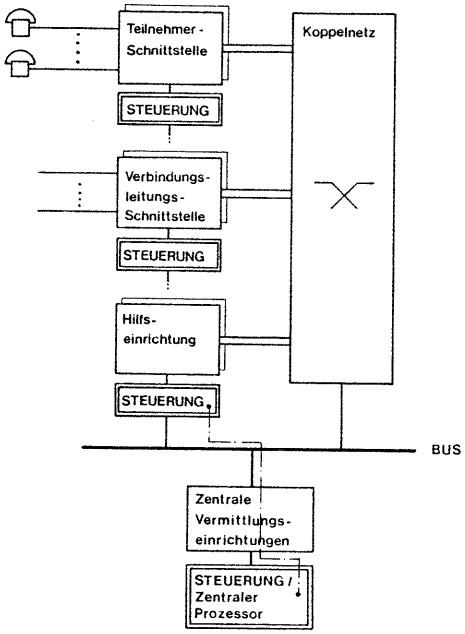
2.3.1 Hardware-Modularisierung und Interprozessor-Kommunikation

In Vermittlungssystemen mit dezentraler oder gemischt zentral/dezentraler Struktur (Kap. 2.2), in denen die Systemintelligenz in mehrere Steuerungseinheiten bzw. Rechner aufgeteilt wird, gewinnt die Interprozessor-Kommunikation eine zentrale Bedeutung bei der Beurteilung der Leistungsfähigkeit des Systems, insbesondere in Überlastsituationen.

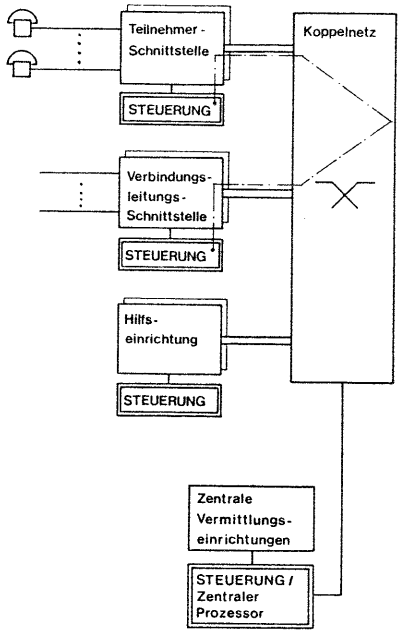
In Bild 2.6 werden zwei Prinzipien der Interprozessor-Kommunikation, die häufig in realen rechnergesteuerten Fernsprechvermittlungssystemen implementiert sind, schematisch dargestellt:

(1) Kommunikation über ein autonomes Bussystem

Der Steuerdatenaustausch erfolgt hier über einen Steuerdatenbus, auf den von allen Steuerungseinheiten gemäß eines Protokolls zugegriffen werden kann. Während bei der dezentralen Struktur mit verteilter Steuerung i.a. alle angeschlossenen Hardware-Module den Steuerdatenbus gleichberechtigt aktivieren können, wird bei der gemischt zentral/dezentralen Struktur der Bus häufig von einer zentralisierten Bus-Steuereinheit aus gesteuert (z.B. dem Ein/Ausgabe-Steuerwerk), die den Ablauf des Steuerdatenaustausches zwischen der zentralen Steuerungseinheit und den peripheren Rechnern kontrolliert.



(1) Kommunikation über ein autonomes Bussystem



(2) Kommunikation über das Koppelnetz

Bild 2.6 Prinzipien der Interprozessor-Kommunikation in rechnergesteuerten Fernsprechvermittlungssystemen.

(2) Kommunikation über das Koppelnetz

Der Steuerdatenaustausch zwischen Steuerungseinheiten geschieht hier über festgeschaltete oder aufgebaute Wege im Koppelnetz. Dieses Prinzip wird bei den dezentralen Strukturen oft angewendet. Abhängig von der Rate der auszutauschenden vermittlungstechnischen Steuerdaten bzw. der Signale zwischen zwei Steuereinheiten werden Wege entweder semipermanent zugeteilt oder nach Bedarf über das Koppelnetz aufgebaut.

2.3.2 Software-Modularisierung und Interprozeß-Kommunikation

In modernen rechnergesteuerten Vermittlungssystemen wird die gesamte System-Software funktionsbezogen in autonome Module aufgeteilt [21-27]. Ähnlich wie bei der modularen Hardware in dezentralen Systemstrukturen erleichtert die modulare Software-Struktur einen schrittweisen Aufbau der System-Software, da die Software modulweise entwickelt, getestet und dokumentiert werden kann. Überdies kann die System-Software aufgrund des modularen Aufbaus und der Schnittstellenfestlegung auf mehrere Steuerungseinheiten bzw. Prozessoren ohne große Anpassungsprobleme verteilt werden.

Eine modular aufgebaute System-Software, bei welcher eine Autonomisierung sowohl der Programme als auch der Datenbereiche realisiert wird, setzt sich aus gleichartig strukturierten, elementaren Software-Bausteinen zusammen: den Software-Maschinen (SMn). Dieses Konzept findet Anwendung in modernen Systemen, insbesondere bei Systemarchitekturen mit verteilter Steuerung [26].

Bild 2.7 zeigt schematisch eine modulare Software-Struktur mit den zugehörigen Software-Maschinen. Die zu einer System-Software gehörenden SMn können auf verschiedene dezentral gesteuerte Rechner nach Bedarf aufgeteilt werden.

Jede SM hat einen Programmteil und einen individuellen Datenbereich. Der Programmteil enthält die für die spezifische SM-Funktion erforderlichen Programme. Die systembezogenen permanenten und semipermanenten Daten sowie die rufbezogenen Daten, die für die Verarbeitung aktiver Teilprozesse und Rufe in der betreffenden SM benötigt werden, sind im Datenteil untergebracht. Um falsche Datenzugriffe zwischen SMn zu vermeiden, dürfen nur die SM-eigenen Programme auf den Datenteil zugreifen [25,26].

Der Datenteil einer SM weist ebenfalls eine modulare Struktur auf und wird unterteilt in rufbezogene Datenblöcke (CCBs: call control blocks). In einem CCB werden alle erforderlichen Daten eines in

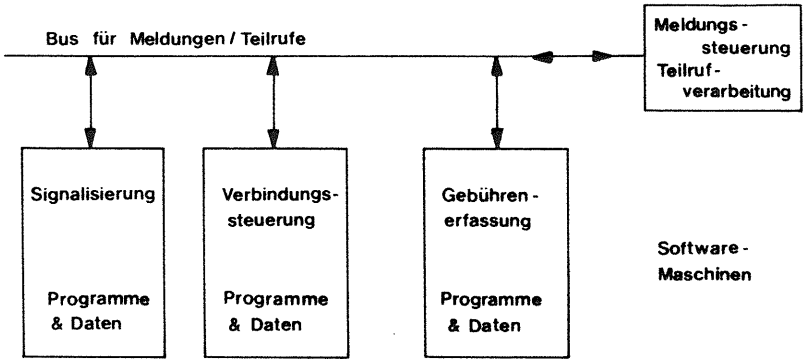


Bild 2.7 *Modulare Software-Struktur in rechnergesteuerten Vermittlungssystemen.*

der betreffenden SM aktiven Rufes gespeichert (z.B. Teilnehmer-Identifikation, Zustand des Rufes, Wählziffern,...).

Die Verarbeitung eines Rufes wird normalerweise von mehreren SMn parallel und/oder seriell gesteuert. Für die Verarbeitung eines Rufes wird in jeder SM ein Teilprozeß aktiviert und ein CCB für die Speicherung rufbezogener Daten reserviert.

Über einen logischen Bus werden Signale in Form von formatierten Meldungen zwischen SMn ausgetauscht, welche die für die Rufverarbeitung benötigte Information enthalten und die logische Zusammenarbeit der in verschiedenen SMn lokalisierten Teilprozesse sicherstellen. Diese Interprozeß-Kommunikation wird noch verstärkt durch den Datenaustausch aufgrund der völligen Trennung der Datenbereiche.

Durch die Einführung modularer Software-Strukturen in rechnergesteuerten Vermittlungssystemen entstehen neue Verkehrsprobleme, die in den Kapiteln 5 und 6 untersucht werden:

- die verkehrsgerechte Dimensionierung der Anzahl von rufbezogenen Datenblöcken (CCBs) in Software-Maschinen (SMn)
- die Prozeßbeschreibung von Teilrufströmen in Form von Meldungen, die zwischen SMn ausgetauscht werden (s.Kap.2.4 und 5)
- die teilrufbezogene Betrachtung von Systemüberlastung und Überlastabwehrmechanismen.

2.4 Verkehrsströme in rechnergesteuerten Vermittlungssystemen

Nachfolgend werden die Verkehrsströme charakterisiert, welche für die in Kap. 5 und 6 diskutierte Modellbildung von Überlastsituationen und Überlastabwehrstrategien von Bedeutung sind.

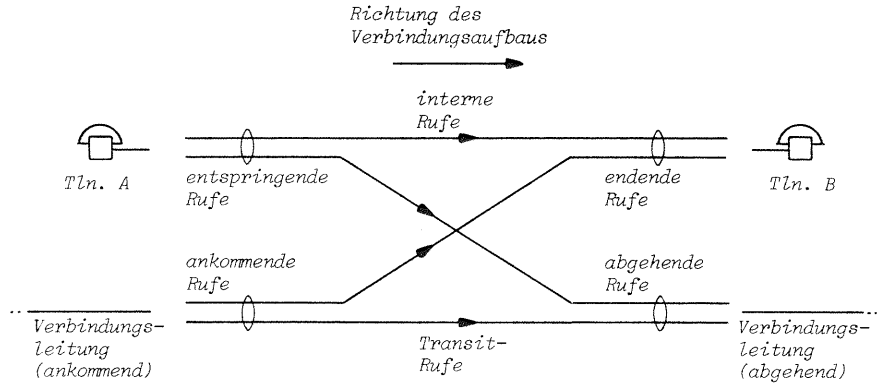


Bild 2.8 *Rufverkehrsströme in Vermittlungssystemen.*

2.4.1 Rufverkehr

Man unterscheidet hinsichtlich der Richtung des Verbindungsaufbaus und des Standortes der an der Verbindung beteiligten Teilnehmer zwischen verschiedenen Typen von Rufen (Verbindungen), welche die Eingangsverkehrsströme eines Vermittlungssystems bilden [12]. Eine Übersicht über diese Rufverkehrsströme gibt Bild 2.8; dabei werden die in einer anderen (fremden) Vermittlungsstelle befindlichen A- bzw. B-Tln. mit den ankommenden bzw. abgehenden Verbindungsleitungen dargestellt.

2.4.2 Teilprozeß- und Teilrufverkehr

Da für die Verarbeitung eines Rufes mehrere Teilprozesse in verschiedenen SMn parallel oder nacheinander aktiviert werden müssen (Bild 2.9), stellen die Anforderungen zur Teilprozeß-Aktivierung für das Vermittlungssystem einen Verkehrsstrom dar, der als Teil-

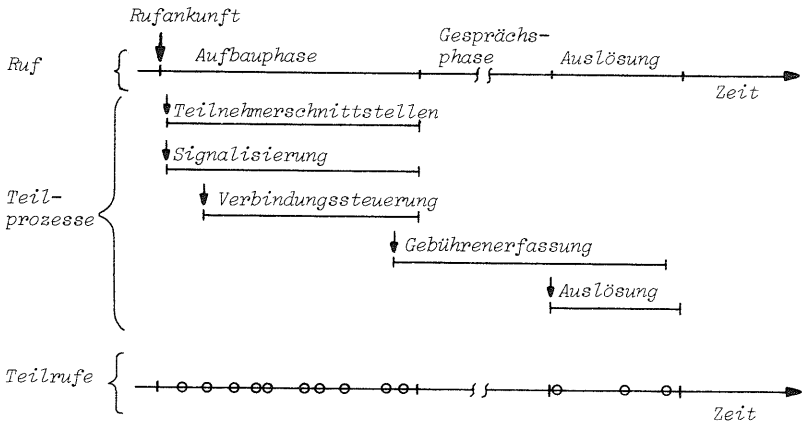


Bild 2.9 *Zusammenhang zwischen Ruf-, Teilprozeß- und Teilrufverkehr.*

prozeßverkehr bezeichnet werden soll. Die Betrachtung des Teilprozeßverkehrs spielt in der Modellbildung zur Dimensionierung von SMn eine bedeutende Rolle (vgl. Kap. 6.2.1).

Die zu einem Ruf gehörenden Teilprozesse erzeugen in ihren aktiven Phasen Steueraufrufe, die von der Steuerungseinheit verarbeitet werden. Sie werden hier als Teilrufe bezeichnet. In Vermittlungssystemen mit verteilter Steuerung und modularer Software-Struktur können alle Meldungen zwischen Software-Maschinen, alle Steueraufrufe in dezentralisierten Steuerungseinheiten sowie formatierte Meldungen zur Interprozessor-Kommunikation als Teilrufe betrachtet werden. Die Modellbildung des Teilrufverkehrs wird in Kap. 5.1 näher erläutert. Kap. 5.2 und 6 stellen teilrufbezogene Modelle vor, die für die Beschreibung von Überlastsituationen sowie für die Leistungsbeurteilung von Überlastabwehrstrategien entwickelt werden.

3. ÜBERLASTPROBLEMATIK IN RECHNERGESTEUERTEN FERNSPRECHVERMITTLUNGSSYSTEMEN

In diesem Kapitel wird die Überlastproblematik in Vermittlungssystemen erläutert. Dabei werden Überlastursachen und -indikatoren aufgezeigt sowie wichtige Überlastabwehrstrategien diskutiert.

3.1 Begriffe

3.1.1 Überlast

Eine vermittlungstechnische Einrichtung wird in der Regel für eine vorgegebene Nennkapazität dimensioniert, d.h. für eine Belastung, welche die Einrichtung bei Gewährleistung der vorgegebenen Verkehrsgüte gerade noch bewältigen kann. Eine Überlast ist eine "über die Nennkapazität hinausgehende Inanspruchnahme einer vermittlungstechnischen Einrichtung" (NTG 0903 Empfehlung 1980, Entwurf [12]).

Ein wesentliches Element muß bei dieser Definition genauer berücksichtigt werden: die zeitliche Entwicklung der Systemüberlastung. Sie beinhaltet eine differenzierte Erfassung der Überlastdynamik, die in Untersuchungen von Überlastabwehrmaßnahmen von grundlegender Bedeutung ist. Einige Aspekte sollen in diesem Zusammenhang erwähnt werden:

(1) Die stochastische Natur des Verkehrsgeschehens in Vermittlungssystemen erschwert die Unterscheidung zwischen einer normalen statistischen Lastschwankung, d.h. einer zufälligen Häufung von Ereignissen bzw. von vermittlungstechnischen Signalen und einer tatsächlichen Überlastsituation. Dieser Sachverhalt muß bei der Festlegung und Dimensionierung von Überlastindikatoren berücksichtigt werden (s. Kap. 3.3).

(2) Die Interaktion zwischen Teilnehmern und dem Vermittlungssystem führt zu komplexen dynamischen Entwicklungen von

Überlastsituationen. Auf eine Senkung der Verkehrsgüte, z.B. bei der Blockierung von Anrufversuchen oder bei langen Wartezeiten, reagieren Teilnehmer mit Rufwiederholungen, die eine zusätzliche, rückkopplungsbedingte Belastung des Systems darstellen. Die Systembelastung bzw. der Ankunftsverkehr dürfen daher nicht getrennt von der Systemreaktion betrachtet werden.

(3) Die Dauer und der zeitliche Verlauf der Überlast müssen für Modelluntersuchungen statistisch hinreichend genau charakterisiert werden. Die Überlastdauer ist grundlegend für die Entscheidung, ob eine Überlastabwehrmaßnahme eingeleitet werden soll oder nicht. Die statistische Beschreibung von Überlast-Verkehrsströmen spielt in Untersuchungen zur Leistungsbeurteilung von Überlastabwehrstrategien eine wichtige Rolle (vgl. Kap. 6.1).

Vom Standpunkt des Fernsprechnetzes aus können Überlastsituationen wie folgt eingeteilt werden:

- Globale Überlastung: Dieser Überlastfall betrifft das gesamte Fernsprechnetz und tritt bei außergewöhnlichen oder verwaltungstechnisch bedingten Ereignissen auf, die starke Erhöhung der Verkehrsintensität zur Folge haben (z.B. Silvester, kurz nach der Tarifumschaltung,...).

- Fokussierte Überlastung: Bei Katastrophenfällen in einem Teil des Fernsprechnetzes folgt i.a. eine Überlastung des betreffenden Vermittlungssystems, da aufgrund von Nachfragen bzw. Mitteilungen der angeschlossenen Fernsprechteilnehmer die abgehenden und ankommenden Anrufraten im System schlagartig ansteigen.

- Lokale Überlastung: Die Überlastsituation beschränkt sich hier auf eine Vermittlungsstelle. Sie wird verursacht z.B. durch Ausfall einer vermittlungstechnischen Einrichtung oder Funktion im betreffenden Vermittlungssystem.

3.1.2 Überlastabwehrstrategie

Nach [12] ist Überlastabwehrstrategie "eine Strategie für die Reaktion einer Vermittlungseinrichtung auf eine Überlastung mit dem Ziel, die Folgen der Überlastung abzuschwächen". In diesem Zusammenhang wird ebenfalls die Verkehrs-Überlastkapazität definiert. Sie ist "die Verkehrsbelastung, die bei vorgegeben reduzierter Verkehrsgüte gerade noch bewältigt wird".

Zur Charakterisierung einer Überlastabwehrstrategie können weitere Eigenschaften herangezogen werden; u.a.:

- Die Wirkbreite: Sie zeigt an, welche vermittlungstechnischen Einrichtungen von der Wirkung der Strategie erfaßt werden. Eine Überlastabwehrmaßnahme kann für einen Teil des Fernsprechnetzes, für ein Vermittlungssystem oder für eine bestimmte Einrichtung (z.B. Rechner, Leitungsbündel, Ein/Ausgabe-Bussystem ...), wirksam sein.
- Die Wirkgeschwindigkeit: Diese Eigenschaft ist entscheidend für die Funktionsfähigkeit einer Überlastabwehrstrategie. Zur Erfassung dieser Eigenschaft sind Messungen an realen Systemen sowie instationäre Untersuchungen in verkehrstheoretischen Modellen erforderlich.

Neben diesen Merkmalen existieren weitere Eigenschaften, die zur Beurteilung von Überlastabwehrstrategien herangezogen werden können. Sie werden in Kap. 3.4 näher betrachtet.

3.2 Überlastursachen

Überlastsituationen können durch systeminterne oder -externe Ursachen ausgelöst werden. Hinsichtlich der Systementwicklung und -dimensionierung sind interne Überlastursachen von primärer Bedeutung. Es handelt sich i.a. um Systemengpässe, welche die Wechselwirkung zwischen Teilnehmerverhalten und Systemreaktion negativ beeinflussen. Die Folge ist eine erhebliche Verringerung der Leistungsfähigkeit des Vermittlungssystems. Überdies kann

die Überlastsituation in andere, noch nicht überlastete Vermittlungssysteme induziert werden.

Eine lokale Überlastsituation kann selten auf eine bestimmte Ursache zurückgeführt werden. Sie ist häufig das Ergebnis einer dynamischen, rückkopplungsbehafteten Entwicklung, an der mehrere Faktoren beteiligt sind. In diesem Zusammenhang werden hier Überlastursachen diskutiert, wobei folgende Hauptfaktoren hinsichtlich einer Klassifizierung herangezogen werden:

- Systemengpässe
- Teilnehmerverhalten
- Induzierung von Überlast im Fernsprechnetz.

3.2.1 Überlastung durch strukturbedingte Engpässe

Anhand zweier Systemarchitekturen, der zentralen Systemstruktur mit konzentrierter Steuerung und der vollkommen dezentralen Systemstruktur mit verteilter Steuerung (vgl. Kap. 2), werden hier strukturspezifische überlastempfindliche Systemmerkmale und -engpässe aufgezeigt.

a) Zentrale Struktur mit konzentrierter Steuerung

Durch die Konzentration aller Steuerungsfunktionen in einem zentralen Rechner stellen die Ein/Ausgabe von vermittlungstechnischen Signalen bzw. von Einstellbefehlen und die Echtzeit-Rechnerkapazität kritische Faktoren hinsichtlich der Überlastung dar.

Tritt bei der Ein/Ausgabe-Funktionseinheit, die den Transfer von Signalen und Befehlen zwischen der Peripherie und der zentralen Steuerungseinheit steuert, ein Engpaß auf, so werden die Durchlaufzeiten von Teilrufen - und damit die Rufverarbeitungszeit - länger. Dies führt zu häufigeren Rufunterbrechungen und -wiederholungen, die eine Überlastung des gesamten Systems zur Folge haben können.

Bedingt durch die Systemstruktur wird der Teilrufverkehr im zentralen Rechner konzentriert. Da die Teilruferzeugung und der Teilrufprozeß hohe statistische Schwankungen aufweisen, können im normalen Betrieb Häufungen von Teilrufen am Rechner auftreten, die zu längeren Wartezeiten oder zum Verlust von Teilrufen führen. Besitzt der zentrale Rechner keine ausreichende redundante Kapazität, so können aus derartigen Lasthäufungen Überlastsituationen entstehen. Dieser Sachverhalt wird in Kap. 6.2.2 näher untersucht.

b) Dezentrale Struktur mit verteilter Steuerung

Durch die Aufteilung der Vermittlungsfunktionen auf mehrere Steuereinheiten wird die Systemzuverlässigkeit erhöht und die Empfindlichkeit des Systems gegen den Ausfall einer Steuerungseinheit hinsichtlich der Überlastbetrachtung geringer. Ein Rechnerausfall ist auf einen Systemteil begrenzt und kann i.a. durch eine redundante Dimensionierung der Anzahl von Steuerungseinheiten aufgefangen werden.

Die dezentralisierte Rufverarbeitung und die Autonomisierung der Software-Funktionen führen jedoch zu neuartigen Systemengpässen, welche die Systemleistung in Überlastsituationen entscheidend beeinflussen:

- Durch die Aufteilung von Vermittlungsfunktionen auf mehrere Rechner erhöht sich der Aufwand für Interprozessor-Kommunikation. Aufgrund der erforderlichen Synchronisation von verteilten Vermittlungsprozessen, die z.B. in verschiedenen Rechnern lokalisiert und für die Verarbeitung eines Rufes zuständig sind, ist die Anzahl von Teilrufen bzw. Steuerungsaufrufen pro Ruf höher als in zentral gesteuerten Systemen. Eine statistische Häufung der zu verarbeitenden Teilrufe kann eine lokale Überlastsituation in einer Steuerungseinheit hervorrufen. Aufgrund der Synchronisation von rufverarbeitenden Teilprozessen wird die Rufverarbeitung in den noch nicht überlasteten Steuerungseinheiten ebenfalls verzögert. Dadurch kann die Überlastsituation in das gesamte System induziert werden.

- Infolge der Autonomisierung der System-Software, bei der eine Trennung der Programm- und Datenbereiche in funktionsbezogene Software-Maschinen (SMn, s. Kap. 2.3.2) realisiert wird, steigt die Anzahl der Steuerungsaufrufe bzw. Meldungen zur Interprozeß-Kommunikation pro Ruf im Vergleich zu nicht-modularen Software-Strukturen. Dies entspricht einer höheren Belastung der jeweiligen Steuerungseinheit.
- Für die Verarbeitung eines Rufes müssen i.a. parallel bzw. nacheinander rufbezogene Datenblöcke (CCBs) in verschiedenen Software-Maschinen bereitgestellt werden. Sind die CCBs in einer SM unzureichend dimensioniert, so können Rufe nach einer Verarbeitungsphase im System blockiert werden (vgl.Kap.6.2.2). Diese sekundäre Rufblockierung entspricht einer ineffektiven Ausnutzung der Systemkapazität, die zur Senkung der Systemleistung und zu Überlastsituationen führen kann.

3.2.2 Überlastung aufgrund des Teilnehmerverhaltens

a) System-Teilnehmer-Interaktion

Die Reaktion von Teilnehmern auf das Verhalten eines Vermittlungssystems ist eine der wichtigen Ursachen von Überlastsituationen. Sie entspricht einem zeitabhängigen Prozeß und ist statistisch nicht einfach zu erfassen. Die Hauptmerkmale des Teilnehmerverhaltens werden nachfolgend erläutert.

Als Folge von statistisch bedingten Systemengpässen werden die Wartezeiten von Teilnehmern in einem Vermittlungssystem länger, z.B. die Wartezeit eines Rufes auf einen Wähltonempfänger, die Wartezeit von Teilrufen auf Verarbeitung,usw.. Abhängig von der Geduld der Teilnehmer ist bei langen Wartezeiten die Wahrscheinlichkeit größer, daß fehlerhafte Reaktionen vorliegen. Ein Wahlbeginn vor dem Wählton führt z.B. zum Empfang einer fehlerhaften bzw. unvollständigen Ziffernfolge; der zugehörige Ruf kann deshalb nicht komplettiert werden. Für den nichtkomplettierten Ruf wird jedoch ein Teil der Systemkapazität aufgewendet (Verarbeitung der Steuerungsaufrufe, Bereitstellung des Ziffernempfängers,

Aktivierung von Teilprozessen, ...), der als ineffektiv zu betrachten ist. Dies verursacht eine Senkung der Leistungsfähigkeit des Vermittlungssystems. Da der Teilnehmer in der Regel den Rufversuch nach einer kurzen Zeit wiederholt, kann ein "Schneeballeffekt" ausgelöst werden: Die Erst- und Wiederholungsrufe stellen eine höhere Systembelastung dar, die wiederum zu längeren Wartezeiten führt, diese erhöhen den Anteil der nichtkomplettierten Rufe und verursachen mehr Rufwiederholungen, usw..

Die hier beschriebene Überlastursache aufgrund der Wechselwirkung zwischen den Teilnehmern und dem Vermittlungssystem kann infolgedessen prinzipiell anhand zweier Wirkmechanismen schematisch dargestellt werden (Bild 3.1):

(1) Wartezeitabhängigkeit: Erhöhte Verkehrsintensität (angebotener Verkehr und Rufwiederholungen) führt zu längeren Wartezeiten und dadurch zu geringerer Rufkomplettierungswahrscheinlichkeit.

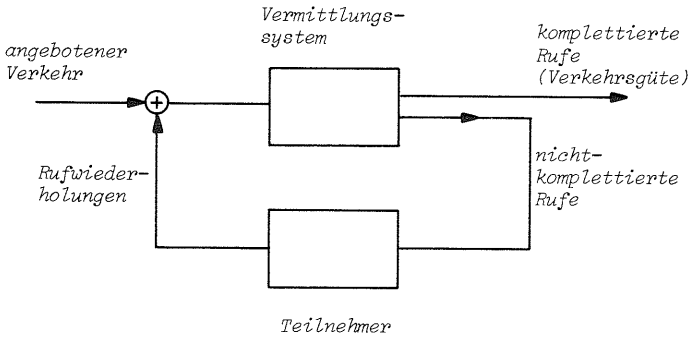


Bild 3.1 *Teilnehmer-System-Interaktion als Überlastursache.*

(2) Rufwiederholung: Nichtkomplettierte Rufe werden nach einer Zeitspanne wiederholt und verursachen eine zusätzliche Systembelastung. Der Rufwiederholungseffekt wird in der Literatur [58-66] mit zwei Parametern charakterisiert: der Wiederholwahrscheinlichkeit und dem Wiederholabstand der Rufe (s. Kap. 5.2). Diese Parameter ändern sich i.a. mit der Anzahl der Wiederholungen, die ein Teilnehmer unternommen hat, abhängig von seiner Geduld-Charakteristik.

b) Saisonale Effekte

Bei außergewöhnlichen Ereignissen (z.B. Silvester) oder anderen, von bestimmten Jahres- bzw. Tageszeiten abhängigen Ereignissen werden Vermittlungssysteme mit einer schlagartigen Anhäufung von Rufversuchen konfrontiert, die weit über der Nennbelastung liegt. Die Systemüberlastung aufgrund saisonaler Effekte erfordert robuste Abwehrstrategien, die das System vor einem Zusammenbruch schützen, einen akzeptablen Durchsatz aufweisen sowie insbesondere die Abfertigung von hochprioritären Rufen (Notfälle) sicherstellen.

3.2.3 Induzierung von Überlast im Fernsprechnetz

Die Problematik der Überlastinduzierung wird im folgenden mit der Darstellung in Bild 3.2, in dem stark vereinfacht ein Ausschnitt des Fernsprechnetzes gezeigt ist, diskutiert. Hierbei ist die prinzipielle Struktur des Netzes zu erkennen: während in den oberen Netzebenen die Vermittlungsstellen maschenförmig untereinander verbunden sind, werden Verbindungswege zwischen Vermittlungsstellen der unteren Netzebenen häufig mit einer sternförmigen Architektur realisiert.

Befindet sich die Vermittlungsstelle B (VSt. B) in einer Überlastsituation, so sind in erster Linie die direkt sternförmig zugeordneten Vermittlungsstellen betroffen. Transitrufe, die von angeschlossenen Teilnehmern über VSt. B abgewickelt werden müssen, werden blockiert und verursachen z.B. durch Rufwiederholungen

eine höhere Belastung dieser Vermittlungsstellen. Durch geeignete Maßnahmen, z.B. die Bereitstellung von Querwegen, können diese Rufe umgelenkt werden (z.B. über VSt.C). Dies erfordert jedoch mehr Steuerungsaufwand im jeweiligen Vermittlungssystem (VSt.b1).

Ein Teil der ankommenden Rufe (vgl. Bild 2.8) wird von der überlasteten Vst.B abgewiesen oder von Teilnehmern aufgrund langer Wartezeit aufgegeben. Dadurch wird die Überlast von VSt.B in benachbarte Vermittlungsstellen induziert, da die für diese Rufe bereits aufgewendete Steuerungskapazität zu keiner Rufkomplettierung führt. Dies verursacht eine Senkung der Rufkomplettierungsrate in den betreffenden Ursprungs-Vermittlungsstellen. Durch den Rufwiederholungseffekt kann bei lang anhaltender Überlastung der VSt.B die induzierte Überlast im Netz weiter verstärkt werden.

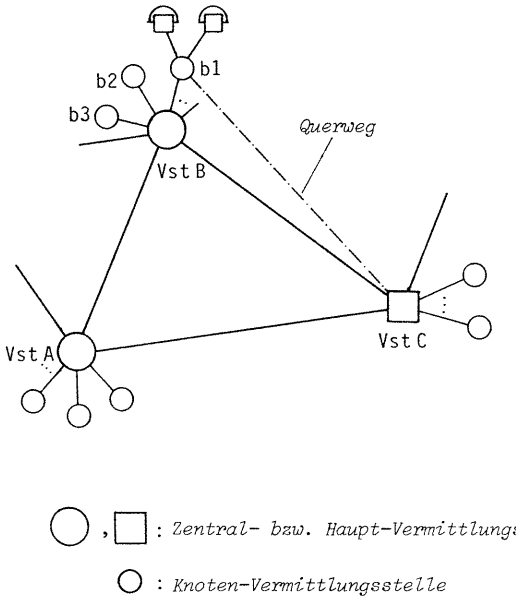


Bild 3.2 Ausschnitt aus einem Fernsprechnetz.

3.3 Überlasterkennung und Überlastindikatoren

Zur rechtzeitigen Erkennung von Überlastsituationen in Vermittlungssystemen werden Indikatoren benötigt [16, 39, 46]. In diesem Unterkapitel werden die Eigenschaften von Überlastindikatoren aufgezeigt sowie eine Klassifizierung der bekannten Indikatoren vorgenommen.

3.3.1 Anforderungen an Überlastindikatoren

Die Wahl der Überlastindikatoren ist ein wesentlicher Schritt zur Entwicklung einer leistungsfähigen Abwehrstrategie. Im allgemeinen haben Überlastindikatoren die Aufgabe, Systemüberlastung räumlich zu lokalisieren, die dynamische Entwicklung der Überlast zu erfassen und den Grad der Überlastgefährdung der betreffenden vermittlungstechnischen Einrichtung anzuzeigen. Demgemäß werden folgende Anforderungen an Überlastindikatoren gestellt:

- Eindeutiger Gültigkeitsbereich: Der Überlastindikator soll genau darüber Auskunft geben, welche Einrichtung, Systemteile und Module überlastet sind. Diese Aussage über die Indikationsbreite ist von Bedeutung, falls mehrere Überlastindikatoren für eine Überlastabwehrmaßnahme gleichzeitig herangezogen werden müssen. Dabei werden durch mehrere Anzeigen der Überlastzustand einer vermittlungstechnischen Einrichtung identifiziert und geeignete Abwehrmaßnahmen gezielt eingeleitet.
- Aktualität: Der Überlastindikator soll den Überlastzustand und die Tendenz zu einer Überlastung rechtzeitig anzeigen. Darüberhinaus soll mit Hilfe desselben Indikators das Ende der Überlastsituation erkannt werden.
- Geringer Steuerungsaufwand: Da ständig ein Teil der Systemkapazität für die Überwachung eines Überlastindikators aufgewendet wird - d.h. auch in Überlastsituationen -, muß dieser Echtzeit-Aufwand möglichst gering gehalten werden.

3.3.2 Typische Überlastindikatoren

Die in der Literatur diskutierten und in realen Systemen benutzten Überlastindikatoren können in zwei Klassen aufgeteilt werden: Indikatoren aus Messungen an vermittlungstechnischen Einrichtungen und Indikatoren aus dem aktuellen Zustand des Vermittlungssystems.

a) Indikatoren aus Messungen

Diese Überlastindikatoren werden aus Messungen und Zählvorgängen an vermittlungstechnischen Einrichtungen gewonnen. Aufgrund der zur Gewinnung dieser Überlastindikatoren erforderlichen Zeitintervalle gelten die dadurch erzeugten Überlastanzeigen nicht für den gegenwärtigen Systemzustand, sondern für die unmittelbare Vergangenheit des Systemgeschehens. Ausgehend von den zuletzt gewonnenen Meßwerten wird die zukünftige Entwicklung des Systems hinsichtlich der Überlastung extrapoliert. Der Aktualitätsgrad dieser Klasse von Indikatoren hängt daher sehr stark von der Geschwindigkeit und der Genauigkeit der Meßmethoden ab. Typische Überlastindikatoren dieser Klasse werden nachfolgend diskutiert:

- Intensität des angebotenen Rufverkehrs: In konstanten Zeitintervallen wird die Anzahl der angebotenen Rufe gemessen. Die Observierung der Rufverkehrsströme kann individuell für entspringende/endende bzw. für ankommende/abgehende Rufe durchgeführt werden (vgl. Bild 2.8). Dadurch erhält man detaillierte Informationen über die Verkehrssituationen im System und die zukünftige Teilruf-Belastung der Steuerungseinheit. Zwei gegensätzliche Aspekte müssen bei der Dimensionierung des Meßintervalls berücksichtigt werden: kurze Meßintervalle bewirken eine aktuelle Überlastanzeige, sie erfordern jedoch einen hohen Verwaltungsaufwand und bedingen stärkere statistische Schwankungen.
- Mittlere Wartezeiten im System: Durch die Messung von Mittelwerten des Wähltonverzuges (dial tone delay) und des Rufverzuges (post-dialling delay) kann der Grad der Systembelastung ermittelt werden. Ebenfalls geben Warte- und Durchlaufzeiten

von Teilrufen sowie Antwortzeiten von vermittlungstechnischen Ereignissen Aufschluß über die momentane Belastung der Steuerungseinheiten im System.

- Belegungs- und Frei-Zeiten der Steuerungseinheiten: Die Rechnerbelegung bzw. -auslastung (processor occupancy) sowie der prozentuale Anteil von freien Zeitabschnitten der Steuerungseinheit zeigen unmittelbar die Intensität der Systembelastung und somit den Grad der Überlastgefährdung an. In Vermittlungsrechnern ist die Durchführung der Messung dieser Überlastindikatoren wegen des Echtzeitbetriebs sehr schwierig und verbraucht überdies einen Teil der Rechnerkapazität, die in Überlastsituation dringend zur Bearbeitung des im System befindlichen Überlast-Verkehrsvolumens benötigt wird.
- Rufblockierungs- und Rufkomplettierungsrate: Die Rate von blockierten Rufen im System zeigt an, in welchem Maße die vermittlungstechnischen Einrichtungen (Koppelnetz, Wahlempfänger, ...) momentan beansprucht sind. Komplementär dazu zeigt die Rate der Rufe, die erfolgreich verarbeitet werden (komplettierte Rufe), die aktuelle Systemleistung an. Diese Indikatoren müssen über mehrere Meßintervalle observiert werden, um eine zuverlässige Aussage über die Überlastgefährdung zu machen.

b) Indikatoren aus dem aktuellen Systemzustand

Diese Überlastindikatoren werden häufig in realen Systemen gebraucht. Es handelt sich hierbei um aktuelle Zustandsgrößen, z.B. die aktuelle Anzahl von aktiven Rufen, Teilprozessen oder auf die Verarbeitung wartenden Teilrufen im System. Diese Überlastindikatoren beschreiben somit den gegenwärtigen Belastungszustand des Vermittlungssystems. Hierbei ist der Verwaltungsaufwand gering, da die benötigten Zustandsinformationen i.a. bei der Durchführung anderer vermittlungstechnischer Funktionen bereits vorhanden sind.

Auf allen Lastebenen - Ruf-, Teilprozeß- und Teilrufebene -

können Systemzustandsinformationen als Überlastindikator genommen werden.

- Rufebene: Charakteristisch für die Systembelastung ist die Anzahl aktiver Ruf im System, insbesondere die in der Aufbauphase befindlichen Rufe, da diese die Intensität des Teilrufverkehrs und die zu erwartende Belastung der Steuerungseinheit bestimmen. Häufig wird diesbezüglich die Anzahl aktiver Wählziffern-Empfänger als Indikator verwendet.
- Teilprozeß-Ebene: Die Anzahl belegter und aktiver rufbezogener Datenblöcke in einem rechnergesteuerten Vermittlungssystem mit einer modularen Software-Struktur entspricht der aktuellen Anzahl von aktivierten Teilprozessen zur Rufverarbeitung. Durch diesen Überlastindikator läßt sich die Intensität von Verkehrsströmen der Teilrufe bzw. Steueraufrufe, die in der jeweiligen Steuerungseinheit verarbeitet und zwischen dezentralen Rechnern ausgetauscht werden, anzeigen.
- Teilruf-Ebene: Die momentane Teilrufbelastung eines Vermittlungsrechners wird z.B. durch die Belegung des Teilruf-Pufferspeichers, d.h. die Anzahl der auf Verarbeitung wartenden Teilrufe, charakterisiert. Die mittlere Wartezeit und somit die Durchlaufzeit von Teilrufen läßt sich ebenfalls mit Hilfe dieses Indikators ermitteln.

3.4 Überlastabwehr und Überlastregelung

3.4.1 Allgemeines

Das Hauptziel der Überlastabwehr in einem Vermittlungssystem ist es, die Komplettierungsrate der akzeptierten Rufe im System zu maximieren (vgl. Definition und Eigenschaft in Kap. 3.1). Mit Hilfe von Überlastabwehrmaßnahmen soll eine Optimierung der Rufkomplettierungsrate in allen Lastsituationen erreicht werden (Bild 3.3).

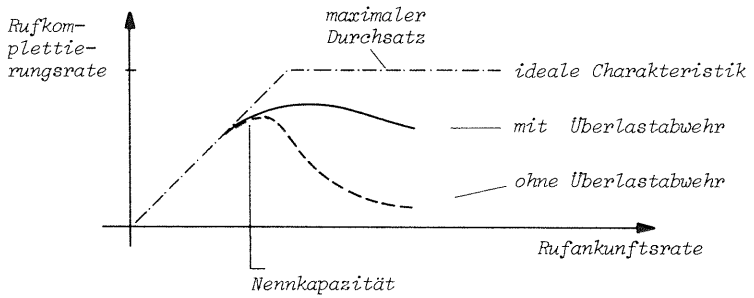


Bild 3.3 *Zielsetzung der Überlastabwehr.*

Bei der Entwicklung und Leistungsbeurteilung von Überlastabwehrmaßnahmen werden folgende Aspekte berücksichtigt [16,46]:

- Die Implementierung einer Überlastabwehrmaßnahme soll einfach sein und möglichst wenig Verwaltungsaufwand verursachen.
- Überlastabwehrmaßnahmen sollen die Verarbeitung der Rufe, die noch komplettierbar sind, nicht negativ beeinflussen. Die Drosselungsmechanismen z.B. sollen sich möglichst nur auf nichtkomplettierbare Rufe, neue Rufe sowie Rufe, die sich in frühen Verarbeitungsphasen befinden, auswirken. Abgewiesene und blockierte Rufe sollen möglichst wenig Systemkapazität in Anspruch nehmen.
- Nach dem Abklingen der Überlastung soll der normale Betrieb des Vermittlungssystems unverzüglich wiederhergestellt werden.
- Bei lokalen Überlastabwehrmaßnahmen müssen Gesichtspunkte zur Verhinderung bzw. Minimierung globaler Überlastinduzierung in Betracht gezogen werden.

3.4.2 Zur Klassifizierung von Überlastabwehrmaßnahmen

Da die meisten Überlastabwehrmaßnahmen systemspezifisch, indikatorabhängig und überdies jeweils auf bestimmten Einrichtungen bzw. Funktionseinheiten eines Vermittlungssystems wirksam sind, ist es nicht einfach, eine überlappungsfreie Klassifizierung vorzunehmen. In diesem Unterkapitel werden Überlastabwehrstrategien, die in der Literatur behandelt bzw. in Kap. 6 vorgestellt und untersucht werden, in folgende Kategorien gegliedert und diskutiert:

- Strukturelle und organisatorische Maßnahmen zur Optimierung der Rufkomplettierung
- Überlastabwehr durch differenzierte Drosselung der Rufannahme
- Globale Überlastabwehr.

a) Strukturelle und organisatorische Maßnahmen zur Optimierung der Rufkomplettierung

Unter Berücksichtigung der Einflüsse des Teilnehmerverhaltens und der strukturbedingten Systemengpässe werden hier Überlastabwehrmaßnahmen vorgestellt, die durch eine optimale Ausnutzung der vorhandenen Systemkapazität oder durch spezielle organisatorische Maßnahmen bei der Rufverarbeitung und der betriebsystemtechnischen Ablaufsteuerung eine Erhöhung der Rufkomplettierungsrate erzielen. Dabei spielt die Abweisung von Rufen bzw. Teilprozessen, die bei den meisten Überlastabwehrstrategien implizit enthalten sind, nur eine untergeordnete Rolle, d.h. der Einfluß des Rufwiederholungseffektes bleibt gering. Zu dieser Klasse von Überlastabwehrmaßnahmen gehören Methoden zur optimalen Ausnutzung der vorhandenen Systemkapazität, Verfahren mit last- und zustandsabhängigen Ablaufsteuerungen sowie spezielle Abwehrstrategien durch wartezeitoptimierende Abfertigungsdisziplinen, die im folgenden erläutert werden. Im einzelnen sind dies:

- Reservierungsmechanismen
- Zwangsauslösung nichtkomplettierbarer Rufe

- Verzögerung niederprioritärer Aufgaben
- Automatische Anpassung der Ablaufsteuerung (Schedule) an die Lastsituation.

Ein Reservierungsmechanismus für rufbezogene Datenblöcke in Software-Maschinen wird in Kap. 6.2.1 vorgestellt und untersucht. Bei dieser Methode wird zur Vermeidung des sekundären Rufblockierungseffektes, der einer ineffektiven Ausnutzung der Systemkapazität entspricht, eine Reservierung von rufbezogenen Datenblöcken in Software-Maschinen, welche Rufe seriell verarbeiten, vorgenommen.

Bei der Überlastabwehrmethode mit Zwangsauslösung nichtkomplettierbarer Rufe (Kap. 6.2.2) wird der Fall betrachtet, daß Rufe, die fehlerhaft verarbeitete Teilrufe generiert haben - d.h. nicht mehr erfolgreich abgewickelt werden können -, weiterhin Teilrufe produzieren, bedingt durch die im Vergleich zu der Teilrufverarbeitung relativ langsam ablaufende Signalisierung. Dadurch wird der Vermittlungsrechner ineffektiv ausgenutzt; die Folge ist eine Verringerung der Rufkomplettierungsrate des Vermittlungssystems. Eine Erhöhung des Durchsatzes komplettierter Rufe wird erreicht, indem Rufe unmittelbar nach einer Fehlreaktion ausgelöst werden - z.B. nach einem blockierten Teilruf oder einem Prozeßsynchronisationsfehler -.

Einige Überlastabwehrmechanismen beruhen auf dem Prinzip, mit Hilfe adaptiver Ablaufsteuerungsverfahren kurzzeitige Lasthäufungen zu glätten und somit mögliche Entwicklungen zu Überlastsituationen zu verhindern. Bei adaptiven Ablaufsteuerungen handelt es sich um Verfahren, die in Überlastsituationen den für die Rufkomplettierung wichtigen Vermittlungsfunktionen (Rufverarbeitung, Signalisierung, Rufauslösung,...) höhere Priorität und anteilmäßig längere Verarbeitungszeit reservieren, während aufschiebbare Vermittlungsaufgaben (Abtastung, Rufannahme, Gebührenerfassung, Routinetests,...) verzögert werden. Zu dieser Gruppe von Überlastabwehrmaßnahmen gehört auch die Verzögerung von Meldungen zur Interprozessor-Kommunikation und von Teilprozeßaktivierungen in

Überlastsituationen. Diese Methode ist jedoch nur wirksam für kurzzeitige Systemüberlastungen.

Zur Optimierung von Durchlaufzeiten der Teilrufe bzw. der Meldungen bei taktgesteuerten Kommunikationsprozeduren werden in [72] einige Mechanismen untersucht, die auf dem Prinzip der systemzustandsabhängigen Unterdrückung von Verwaltungszeiten basieren. Dabei wird der in [71] behandelte Grundmechanismus der taktgesteuerten Übernahme vermittlungstechnischer Ereignisse zugrundegelegt. Abhängig von dem aktuellen Systemzustand, z.B. von der Anzahl wartender Teilrufe im sendenden bzw. empfangenden Prozessor, wird zum Taktzeitpunkt festgelegt, ob die bevorstehende Kommunikationsprozedur zur Teilrufübergabe ausgeführt oder unterdrückt wird. Dadurch werden eine Optimierung der Ausnutzung der Prozessorkapazität zur Teilrufverarbeitung und eine Minimierung der Wartezeiten von Teilrufen erzielt.

Die Wartezeitcharakteristik von Teilnehmern wird bei der Überlastabwehrstrategie in [44] berücksichtigt ("1E6 generic" für ESS1-Vermittlungssysteme). Da die Rufaufgabewahrscheinlichkeit mit der Wartedauer von Teilnehmern ansteigt, wird hier die Abfertigungsdisziplin LIFO (last-in first-out) für die Rufannahme realisiert. Anrufversuche von Teilnehmern, deren Wartezeit eine vorgegebene Schwelle (z.B. 20-30sec) überschritten hat und die damit als geduldige Teilnehmer betrachtet werden - zugehörige Rufe sind mit großer Wahrscheinlichkeit komplettierbar - werden als neue Versuche eingestuft, die aufgrund der LIFO-Disziplin erneut vorrangig behandelt werden. Durch diese Überlastabwehrmethode wird eine Optimierung des Durchsatzes erreicht, da auch bei starken Überlastungen eine Anzahl von Teilnehmern existiert, die eine kurze Wartezeit erfahren.

b) Differenzierte Drosselung der Rufannahme

Es handelt sich hierbei um Überlastabwehrmaßnahmen, bei denen
- zur Verhinderung einer bevorstehenden Überlastentwicklung oder zum Abbau bestehender Überlastsituationen im System - die ankommenden Verkehrsströme gezielt gedrosselt werden:

- Graduelle, systemzustandsabhängige Rufblockierung
- Hysteresebehaftete Rufblockierung.

Eine einfache Rufblockierungsstrategie besteht darin, abhängig vom Belastungszustand des Vermittlungssystems einfallende Rufversuche graduell abzuweisen (vgl. Kap. 6.1.2). Dieses Verfahren kann z.B. durch eine reduzierte Abtastung - indem Teile des Anschlußbereiches vom Abtastvorgang kurzzeitig ausgeschlossen werden - oder eine Verlangsamung des Abtastzyklus realisiert werden. Als Überlastindikator kann die Anzahl aktiver Rufe oder aktivierter Teilprozesse oder eine Warteschlangenlänge im System dienen.

Eine Variante der graduellen Rufblockierung ist die Begrenzung der Anzahl von Rufen, die in einem Zeitintervall angenommen werden dürfen. In [45] wird eine adaptive Annahmestrategie vorgestellt, die auf einem Prädiktionsverfahren des Verkehrsverhaltens basiert. Die in [40] diskutierte Rufannahmemethode berechnet die Anzahl der in einem Zeitintervall akzeptierbaren Rufe aus dem Grad der Prozessorbelegung und den Verkehrswerten der vorausgegangenen Zeitintervalle. [34] behandelt eine Überlastabwehrstrategie, bei der eine zustandsabhängige, stufenweise Sperrung von Wählziffernempfängern implementiert wurde, wodurch die Anzahl der sich in der Aufbauphase befindenden Rufe begrenzt wird.

Weist der verwendete Überlastindikator starke statistische Schwankungen auf, so ist es zweckmäßig, die Rufannahmestrategie mit einer hysteresebehafteten Charakteristik zu versehen. Da das Schwingen zwischen dem aktivierten und dem deaktivierten Zustand der Überlastabwehr dadurch vermieden wird, läßt sich der für die Überlastabwehrmaßnahme benötigte Verwaltungsaufwand reduzieren. In Kap. 6.1.1 wird eine von der Anzahl aktiver Teilprozesse gesteuerte Rufannahmestrategie mit einer Zweipunkt-Regelung vorgestellt und untersucht.

Der Belastungszustand des Systems wird präziser erfaßt, wenn Über-

lastindikatoren aus den drei Lastebenen - Ruf-, Teilprozeß- und Teilrufebene - zusammen betrachtet werden. Eine Rufblockierung erfolgt z.B. dann, wenn die Schwellenwerte der Anzahl der aktiven Rufe und der auf Verarbeitung wartenden Teilrufe überschritten sind (Kap. 6.2.2g).

c) Globale Überlastabwehr

Zu dieser Klasse gehören Überlastabwehrmaßnahmen, welche die globalen Gesichtspunkte der Überlastregelung berücksichtigen [41,42]. Mit diesen Methoden soll der Durchsatz im gesamten Fernsprechnetz optimiert werden, wobei bestehende lokale Überlastsituationen in den einzelnen Vermittlungsstellen nicht erheblich weiter verschärft werden. Im einzelnen sind dies:

- Priorisierung ankommender bzw. endender Verkehrsströme
- Adaptive Verkehrslenkung.

Eine einfache Maßnahme besteht darin, den ankommenden und endenden Verkehrsströmen (vgl. Bild 2.8) höhere Priorität bei der Rufannahme und -verarbeitung einzuräumen. Die Blockierungswahrscheinlichkeit der Rufe, die bereits von anderen Vermittlungsstellen bzw. von anderen Vermittlungseinrichtungen im betrachteten System erfolgreich verarbeitet werden, wird geringer. Dadurch lassen sich eine Reduzierung ineffektiver Verarbeitungszeiten und eine Erhöhung des Durchsatzes im Fernsprechnetz erreichen.

Geeignet zur Regelung der fokussierten und globalen Überlastung ist die adaptive Verkehrslenkung. Dabei muß die Zustandsinformation des gesamten Netzes - bzw. eines großen Bereichs des Netzes - in einer sog. Netzkontrollstelle verfügbar sein. Anhand dieser aktuellen Information wird der Verkehr um die überlasteten Vermittlungsstellen weiträumig umgeleitet.

Bei einer periodisch auftretenden Überlastung einer Vermittlungsstelle oder eines Bereiches im Netz - z.B. zu einer bestimmten Tageszeit - läßt sich die Überlastsituation durch eine temporäre,

fest vorgeplante Änderung der Verkehrslenkungstabellen in den benachbarten Vermittlungsstellen entschärfen.

Die hier angesprochenen Verfahren gewinnen in den modernen Nachrichtennetzen zunehmend an Bedeutung und werden unter dem Begriff der "Netzführung" (network management) zusammengefaßt.

4. METHODEN ZUR MODELLANALYSE

Das Ziel dieses Kapitels ist es, die Methodik der Modellbildung zu erläutern und einen Überblick über die in dieser Arbeit angewandten Methoden zur Modellanalyse zu geben.

4.1 Zur verkehrstheoretischen Modellbildung

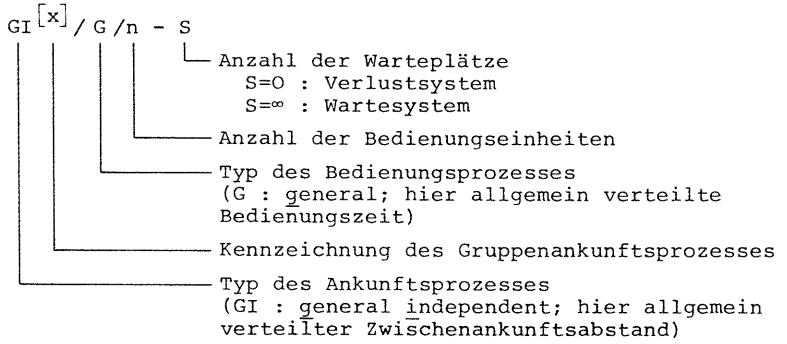
Für die Untersuchung komplexer Vorgänge in realen Systemen werden häufig Modelle benötigt. In Nachrichtensystemen, z. B. Fernsprechvermittlungssystemen, sind es verkehrstheoretische Modelle, die eine quantitative und qualitative Beschreibung des Verkehrsgeschehens sowie eine Beurteilung der Systemreaktion erlauben.

Ein verkehrstheoretisches Modell beschreibt das Ablaufgeschehen in allen seinen zeitlichen und logischen Zusammenhängen mit Hilfe weniger abstrakter Modellelemente, welche die realen Systemkomponenten und das logische Zusammenspiel zwischen ihnen widerspiegeln. Die Modellbildung beinhaltet infolgedessen das Abbilden der Systemorganisation bzw. des dynamischen Systemgeschehens auf entsprechende Modellkomponenten und modellbezogene äquivalente Vorgänge.

Zur Untersuchung des Überlastproblems werden in den Kapiteln 5 und 6 Verkehrsmodelle entwickelt und untersucht, welche

- das breite Spektrum der Überlast- und Überlastabwehrprobleme in abgrenzbaren Teilproblemen behandeln. Dabei sind die wesentlichen Systemparameter und ihre gegenseitige Beeinflussung im Modell enthalten.
- die Festlegung und Dimensionierung der Systemparameter sowie die Leistungsbeurteilung der zu untersuchenden Überlastabwehr-Strategien erlauben.

Für die Modellanalyse werden exakte und simulative Untersuchungsmethoden angewendet, die in den Unterkapiteln 4.2 und 4.3 behandelt werden. Die Kennzeichnung der Modelle erfolgt mittels der erweiterten Form der von Kendall (1954) eingeführten Notation :



Zur Erläuterung des Verkehrsgeschehens in einem Verkehrsmodell werden in Bild 4.1 die Struktur und der Zustandsprozeß des Warteverlustsystems GI/G/1-S skizziert. Das System besteht aus einer Bedienungseinheit und einer Warteschlange mit begrenzter Kapazität (S = 3). Anforderungen, die zu den Zeitpunkten $t_{A,i}$ eintreffen, werden von einem Ankunftsprozeß erzeugt, bei dem die Zeitintervalle $(t_{A,i}-t_{A,i-1})$ unabhängig voneinander und mit einer Verteilungsfunktion (GI) charakterisierbar sind. Die Bedienungsdauern der Anforderungen sind hier ebenfalls mit einer Verteilungsfunktion (G) beschreibbar.

Betrachtet werde nun eine Test-Anforderung, die zum Zeitpunkt $t_{A,n}$ eintrifft, in dem die Bedienungseinheit belegt ist und alle Warteplätze verfügbar sind $(X(t_{A,n}^-)=1)$. Die Test-Anforderung wartet in der Warteschlange bis zum Freiwerden der Bedienungseinheit. Bei jedem Bedienungsende zum jeweiligen Zeitpunkt $t_{E,i}$ wird eine wartende Anforderung entsprechend einer festgelegten Abfertigungsdisziplin zur Bedienung übernommen. Bei der in diesem Beispiel angegebenen Abfertigungsdisziplin FIFO (first-in first-out) wird die Test-Anforderung zum Zeitpunkt $t_{E,n}$ in die Bedienungseinheit aufgenommen. Unmittelbar nach der Bedienung verläßt die Test-Anforderung das System $(t_{E,n+1})$.

Eine Anforderung wird abgewiesen, wenn sie zum Ankunftszeitpunkt das System im vollbelegten Zustand vorfindet $(t_{A,n+4})$.

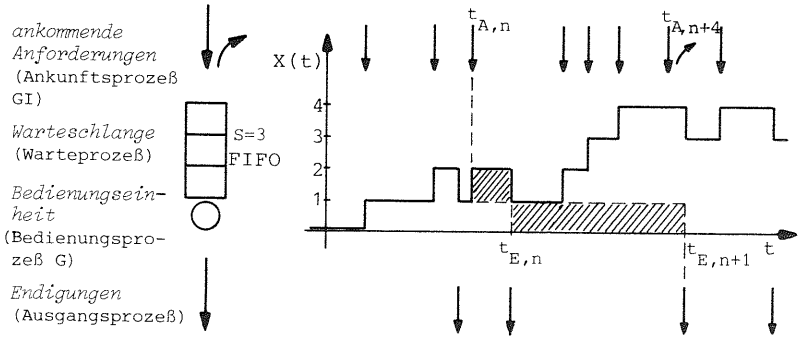


Bild 4.1 *Struktur und Zustandsprozeß eines Warteverlustsystems GI/G/1-S (S=3).*

Einige komplexere Modelle zur Beschreibung von Überlastproblemen, die in Kap. 5 und 6 vorgestellt werden, können jedoch nicht vollständig und eindeutig mit der Kendall'schen Notation charakterisiert werden. Es handelt sich dabei um Modelle, bei denen

- verschiedene Lastebenen parallel zu beschreiben sind (z.B. Ruf- und Teilrufströme, s. Kap. 6.2.2).
- Rückkopplungen zur Beschreibung von Überlastabwehrstrategien vorhanden sind.
- Abhängigkeiten zwischen Modellkomponenten berücksichtigt werden, z.B. für die Untersuchung des Teilnehmerverhaltens.
- instationäre Verkehrsströme zur Überlastbeschreibung einbezogen werden müssen.

Die in diesem Zusammenhang entstandenen Modellkomponenten erfordern neben den herkömmlichen Methoden z. T. neue analytische und simulative Ansätze zur Modellanalyse.

4.2 Analytische Methoden zur Modelluntersuchung

In diesem Kapitel werden analytische Methoden für die Untersuchung verkehrstheoretischer Modelle behandelt. Die Darstellung beschränkt sich hier in erster Linie auf die Analysemethoden, die in dieser Arbeit angewendet werden. Das umfangreiche Spektrum der analytischen Verfahren zur Modelluntersuchung wird in den Standardwerken der Bedienungstheorie (Verkehrstheorie, Warteschlangentheorie) [2, 3, 4, 7, 8] ausführlich dargestellt.

4.2.1 Analyse Markoff'scher Prozesse

a) Definition

Die Markoff'schen Prozesse bilden eine Klasse stochastischer Prozesse, die eine wesentliche Rolle in der Verkehrstheorie spielen. Ein stochastischer Prozeß X(t) heißt Markoff'scher Prozeß, wenn seine zukünftige Entwicklung nur vom gegenwärtigen Prozeßzustand abhängig ist. Ist x_i der Zustand des Prozesses X(t) zum Zeitpunkt t_i ($t_{i-1} < t_i < t_{i+1}$, i=1,2,..), so kann die Definition des Markoff'schen Prozesses X(t) folgendermaßen formuliert werden :

$$P\{X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n, \ldots, X(t_0) = x_0\}$$

$$= P\{X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n\} \quad . \qquad (4.1)$$

Betrachtet man den Prozeß zum Zeitpunkt t_n, so ist die ganze Entwicklung des Prozesses in der Vergangenheit vollständig im Prozeßzustand x_n enthalten. Dieser Sachverhalt wird als die Eigenschaft der Gedächtnislosigkeit der Markoff'schen Prozesse (oder Markoff-Eigenschaft) bezeichnet. Das Vorhandensein der Markoff-Eigenschaft kann auf einen Ankunftsprozeß, einen Bedienungsprozeß oder einen Zustandsprozeß bezogen werden.

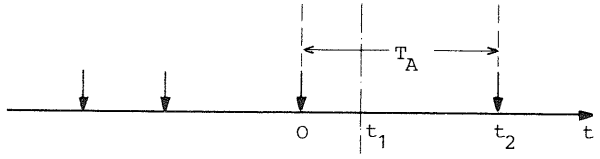


Bild 4.2 *Zur gedächtnislosen Eigenschaft Markoff'scher Prozesse.*

Bild 4.2 zeigt einen Ankunftsprozeß, welcher zum Zeitpunkt t_1 betrachtet wird. Die letzte Anforderung sei nun zum Zeitpunkt $t = 0$ eingetroffen und T_A sei die Zufallsvariable für die Zwischenankunftszeit (Zwischenankunftsabstand oder kurz Ankunftsabstand).

Die Markoff'sche Eigenschaft für den Ankunftsprozeß kann wie folgt ausgedrückt werden :

$$P\{T_A > t_2 \mid T_A > t_1\} = P\{T_A > t_2 - t_1\}. \tag{4.2}$$

Die Beziehung (4.2) besagt, daß die Zeitspanne bis zur Ankunft der nächsten Anforderung nicht davon abhängt, wie lange der letzte Ankunftszeitpunkt zurückliegt. Es kann gezeigt werden, daß die negativ exponentielle Verteilung die einzige Verteilung ist, welche die Eigenschaft (4.2) aufweist:

$$F(t) = P\{T_A \leq t\} = 1 - e^{-\lambda t}, \tag{4.3}$$

mit $E[T_A] = 1/\lambda$.

b) Die Kolmogoroff'schen Gleichungen

Gegeben sei der Systemzustandsprozeß $\{X(t),\ t \geq 0\}$, welcher die Markoff'sche Eigenschaft besitzt. Die Übergangswahrscheinlichkeit

$$p_{ij}(u) = P\{X(t+u) = j \mid X(t) = i\}, \quad u \geq 0, \tag{4.4}$$

beschreibt die Wahrscheinlichkeit für den Zustandsübergang "von i nach j" in der Zeitspanne u. Gl.(4.4) setzt voraus, daß der Prozeß homogen ist, d.h. $p_{ij}(u)$ nicht vom Zeitpunkt t abhängt. Betrachtet man nun die Prozeßentwicklung während zweier aufeinanderfolgender Zeitabschnitte t und u, so gilt die Chapman-Kolmogoroff-Gleichung [8,10]

$$p_{ij}(t+u) = \sum_k p_{ik}(t)\ p_{kj}(u) \ . \tag{4.5}$$

Der Übergang $i \to j$ setzt sich aus zwei Übergängen zusammen: $i \to k$ in der Zeitspanne t und $k \to j$ in der Zeitspanne u, wobei k beliebig sein kann. Die *Kolmogoroff'sche Vorwärtsgleichung* der Übergangswahrscheinlichkeiten erhält man aus (4.5) durch den Grenzübergang $u \to 0$

$$\frac{d}{dt} p_{ij}(t) = -q_j p_{ij}(t) + \sum_{k \neq j} q_{kj} p_{ik}(t) \ , \tag{4.6}$$

mit

$$q_{ik} = \lim_{u \to 0} \frac{p_{ik}(u)}{u} \ ; \ q_j = \lim_{u \to 0} \frac{1 - p_{jj}(u)}{u} \ .$$

Parallel zur Gl. (4.6) existiert eine - hier nicht behandelte - Kolmogoroff'sche Rückwärtsgleichung, die zur Bestimmung von Wartezeitverteilungsfunktionen häufig verwendet wird. q_{ik} und q_j werden Übergangswahrscheinlichkeitsdichten oder *Übergangsraten* genannt. Man erkennt die Markoff'sche Eigenschaft darin, daß die Übergangsraten nur vom gegenwärtigen Zustand abhängig sind und nicht von der Prozeßvergangenheit.

Der betrachtete Prozeß $\{X(t),\ t \geq 0\}$ beschreibt den Zustand eines Systems mit den Zustandswahrscheinlichkeiten

$$P_j(t) = P\{X(t) = j\} \ , \ j = 0,1,\ldots$$

und der Anfangsverteilung zum Zeitpunkt $t = 0$

$$P_j(0) \ ; \ j = 0,1,\ldots \ .$$

Die Zustandswahrscheinlichkeiten zu einem beliebigen Zeitpunkt t lassen sich aus der Anfangsverteilung mit Hilfe des Gesetzes der totalen Wahrscheinlichkeit berechnen

$$P_j(t) = \sum_i P_i(0)\ p_{ij}(t) \ . \tag{4.7}$$

Durch Einsetzen von Gl. (4.6) in Gl. (4.7) ergibt sich die Kolmogoroff'sche Vorwärtsgleichung der Zustandswahrscheinlichkeiten:

$$\frac{d}{dt} P_j(t) = -q_j P_j(t) + \sum_{k \neq j} q_{kj} P_k(t) \; . \qquad (4.8a)$$

Dabei gilt die Normierungsbedingung

$$\sum_j P_j(t) = 1 \; . \qquad (4.8b)$$

Gemäß Gl. (4.6) ist q_{kj} die Übergangsrate vom Zustand k nach Zustand j, q_j die Übergangswahrscheinlichkeitsdichte für das Verlassen des Zustandes j. (4.8a,b) sind die Grundgleichungen für analytische Untersuchungen des instationären Verhaltens Markoff'scher Systeme. In Kap. 5.2.2 findet sich ein Beispiel zur instationären Systemanalyse.

In den meisten Anwendungen ist das Systemverhalten im stationären Falle von großer Bedeutung. Dabei betrachtet man das System im eingeschwungenen Zustand:

$$P_j = P_j(\infty) = \lim_{t \to \infty} P_j(t) \; ,$$

$$\lim_{t \to \infty} \frac{d}{dt} P_j(t) = 0 \; . \qquad ...(4.9)$$

Aus (4.8) und (4.9) folgt die Kolmogoroff'sche Gleichung für den stationären Fall

$$q_j P_j = \sum_{k \neq j} q_{kj} P_k \; , \quad j = 0,1,... \; . \qquad (4.10)$$

Gl. (4.10) besagt, daß während eines infinitesimal kurzen Zeitabschnitts dt die Wahrscheinlichkeit für das Verlassen des Zustandes j gleich der Wahrscheinlichkeit für das Erreichen dieses Zustandes aus allen anderen Zuständen ist. Der Prozeß befindet sich dabei im statistischen Gleichgewicht.

Durch die Auflösung des linearen Gleichungssystems gemäß Gl.(4.10) können die Zustandswahrscheinlichkeiten bestimmt werden. Dies kann explizit mit einem rekursiven Algorithmus oder durch ein iteratives Verfahren erfolgen, abhängig vom System und dessen

zugehörigen Übergangsraten. In Kap. 6.1.1 findet sich ein Beispiel für eine explizite Lösung mit Hilfe einer Rekursionsformel, während in Kap. 5.2.1 ein rekursiver Algorithmus zur numerischen Bestimmung der Zustandswahrscheinlichkeiten angewendet wird. Die Iterationsmethode wird bei der Analyse einiger Modelle in dieser Arbeit benutzt. Die Kolmogoroff'sche Gleichung (4.10) gilt i.a. für eine beliebige Ansammlung von Zuständen, welche auch Makrozustand genannt wird. Die Betrachtung von Makrozuständen wird z.B. in Kap. 5.2.1 zur Herleitung von Zustandsgleichungen hinzugezogen.

4.2.2 Methode der eingebetteten Markoff-Kette

Die Markoff'sche Eigenschaft besitzen die Prozesse in Verkehrsmodellen, bei denen der Ankunftsprozeß und der Bedienungsprozeß einem Poisson-Strom bzw. einer negativ exponentiellen Verteilung (Gl. 4.3) entsprechen. Falls *eines* dieser Modellelemente gedächtnisbehaftet ist, z.B. indem es eine allgemeine Verteilung aufweist, ist der Systemzustandsprozeß nicht gedächtnisfrei. Man kann jedoch häufig einzelne Zeitpunkte des Prozesses finden, welche die Markoff'sche Eigenschaft besitzen. In diesen sog. *Regenerationspunkten* verliert der Prozeß sein Gedächtnis, z.B. zum Ankunftszeitpunkt eines allgemein verteilten Ankunftsprozesses oder am Ende einer allgemein verteilten Bedienungszeit. Die Prozeßentwicklung ist von dort aus unabhängig von der Vergangenheit.

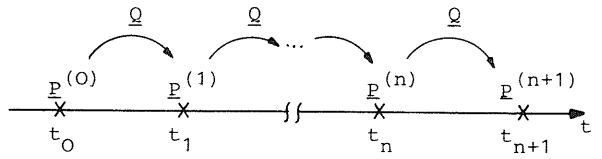


Bild 4.3 *Zustandswahrscheinlichkeit der eingebetteten Markoff-Kette.*
X : *Regenerationspunkte*

Die Regenerationspunkte t_i (i=0,1,..,n,..) bilden die sog. eingebettete Markoff-Kette (Bild 4.3). Der Zustandsprozeß X(t) besitze zum Zeitpunkt t_i den Wahrscheinlichkeitsvektor

$$\underline{P}^{(i)} = \begin{bmatrix} p_0^{(i)} \\ p_1^{(i)} \\ . \\ . \end{bmatrix} \quad ; \; i = 0,1,\ldots , \qquad (4.11)$$

wobei

$$p_k^{(i)} = P\{X(t_i) = k\} \quad ; \; k=0,1,\ldots .$$

Die Beziehung zwischen den Wahrscheinlichkeitsvektoren zweier beliebiger aufeinanderfolgender Regenerationspunkte t_n und t_{n+1} wird mit einer Übergangswahrscheinlichkeitsmatrix $\underline{Q}$ hergestellt :

$$\underline{Q} = \begin{bmatrix} q_{00} & q_{10} & \cdots \\ q_{01} & q_{11} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} ,$$

wobei

$$q_{jk} = P\{X(t_{n+1})=k \mid X(t_n)=j\} \quad ; \; k,j = 0,1,2,.. . \qquad (4.12)$$

Der Wahrscheinlichkeitsvektor $\underline{P}^{(n+1)}$ kann berechnet werden zu

$$\underline{P}^{(n+1)} = \underline{Q} \cdot \underline{P}^{(n)} . \qquad (4.13)$$

Im stationären Zustand ist der Wahrscheinlichkeitsvektor unabhängig vom Zeitindex i

$$\underline{P}^{(n+1)} = \underline{P}^{(n)} = \underline{P} . \qquad (4.14)$$

Aus Gl.(4.13) und Gl.(4.14) folgt die Bestimmungsgleichung für den Zustandswahrscheinlichkeitsvektor der Markoff-Kette im stationären Falle :

$$\underline{P} = \underline{Q} \cdot \underline{P} . \qquad (4.15)$$

Aus Gl.(4.15) ist ersichtlich, daß sich die Zustandswahrscheinlichkeiten der eingebetteten Markoff-Kette aus der Eigenvektorbestimmung der Übergangswahrscheinlichkeitsmatrix $\underline{Q}$ ergeben. Aus den gewonnenen Wahrscheinlichkeiten lassen sich modellspezifische Charakteristiken berechnen.

4.2.3 Die Little'sche Formel

Eine häufig benutzte Beziehung zur Berechnung von mittleren Wartezeiten in Warteschlangensystemen wird von Little [73] angegeben.

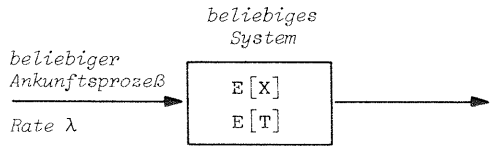


Bild 4.4 Mittelwertbetrachtung nach Little.

Betrachtet werde ein *System*, das eine beliebig herausgegriffene Teilmenge eines Warteschlangenmodells darstellt. Ein beliebiger Ankunftsprozeß mit der mittleren Ankunftsrate λ bildet den Eingangsverkehr. Im Mittel halten sich E[X] Anforderungen mit der mittleren Aufenthaltszeit E[T] auf. Dann gilt die Little'sche Formel :

$$\lambda . E[T] = E[X] . \qquad (4.16)$$

4.3 Simulative Methoden zur Modelluntersuchung

In Untersuchungen verkehrstheoretischer Modelle werden häufig Simulationsstudien benötigt, welche

- die Validierung von exakt oder approximativ gewonnenen analytischen Ergebnissen ermöglichen
- quantitative und qualitative Aussagen über Problemstellungen erlauben, deren Komplexität keiner analytischen Lösungsmethode zugänglich ist.

In diesem Unterkapitel wird zunächst das Prinzip der oft verwendeten Methode der zeittreuen Simulation erläutert [3, 11]. Anschließend wird die Vorwärts-Integral-Methode vorgestellt; sie wurde in dieser Arbeit entwickelt und erlaubt eine zeittreue Simulation verallgemeinerter Poisson-Prozesse zur instationären Modelluntersuchung. Mit Hilfe der Vorwärts-Integral-Methode können das dynamische Systemverhalten und die Systemantwort auf Poisson-Verkehre mit beliebigen zeitlichen Intensitätsschwankungen untersucht werden [56].

4.3.1 Allgemeines über die zeittreue Simulation

Bei der zeittreuen, ereignisgesteuerten Simulation wird das Ablaufgeschehen im System durch eine Folge von Ereignissen abgebildet. Die Systemstruktur und die Betriebsarten sowie die Reihenfolge der Ereignisverarbeitung werden durch ein Programm auf einem Digitalrechner abgebildet. Das Programm zur zeittreuen Simulation hat im wesentlichen zwei Aufgaben: die Abbildung des Verkehrsmodells auf entsprechende Programmkomponenten, die Erzeugung und Abarbeitung von Ereignissen und die Durchführung von statistischen Messungen.

Das programmäßige Abbilden des Verkehrsmodells beinhaltet die Anordnung und die Verwaltung von modellbezogenen Datenstrukturen. In diesem Datenbereich sind die aktuellen Zustände von Modellkomponenten gespeichert. Trifft ein Ereignis ein, das eine Zustandsänderung zur Folge hat, so wird die Änderung in das entsprechende Datenteil eingetragen.

Zur Vorplanung der Ereignisse wird ein "Ereignis-Kalender" geführt, der alle für die nächste Zukunft des Prozesses ermittelten Ereignisse enthält. Zu diesem Zweck benötigt man eine "Systemzeit", die am Anfang eines Simulationslaufes initialisiert wird. Das Simulationsprogramm bearbeitet nacheinander die Ereignisse in dem Kalender. Nach der Verarbeitung eines Ereignisses wird der nächste Ereigniszeitpunkt ermittelt. Das Programm springt demgemäß von einem Ereigniszeitpunkt zum nächsten; die Zeitspanne dazwischen wird simuliert.

Für die Bestimmung zukünftiger Ereigniszeitpunkte benötigt man zufallsbehaftete Zeitintervalle (z.B. Ankunftsabstände, Bedienungsdauern), wobei die Verteilungsfunktionen bekannt sind. Zur Ermittlung dieser statistisch verteilten Zeitintervalle geht man in der Regel von einem Zufallsgenerator aus, der gleichverteilte Pseudo-Zufallszahlen im Intervall (0,1) erzeugt. Die ausgewürfelte Zufallszahl z wird zu einem Funktionswert der vorgegebenen Verteilungsfunktion transformiert (z.B. $F_A(t)$ in Bild 4.5).

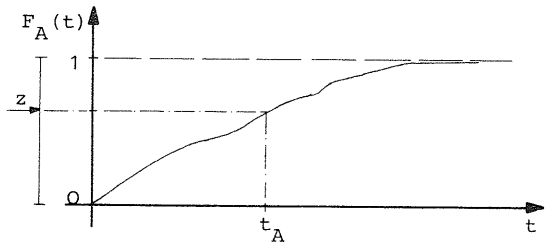


Bild 4.5 *Ermittlung eines zufallsabhängigen Zeitintervalls.*

z : *ausgewürfelte gleichverteilte Pseudo-Zufallszahl.*

t_A : *ermitteltes Zeitintervall mit der Verteilungsfunktion* $F_A(t)$.

Aufgrund der Gleichverteilung der Zufallszahl z kann statt der Verteilungsfunktion $F_A(t)$ die komplementäre Verteilungsfunktion $F_A^c(t) = 1-F_A(t)$ für die Transformation in Bild 4.5 verwendet werden. Anhand der negativ exponentiellen Verteilung wird nachfolgend die Ermittlung des zufallsabhängigen Zeitintervalls t_A exemplarisch gezeigt. Mit:

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

oder

$$F_A^c(t) = e^{-\lambda t}$$

erhält man nach der Transformation der Zufallsvariablen:

$$z = F_A^c(t_A) = e^{-\lambda t_A}$$

oder

$$t_A = -\frac{1}{\lambda} \ln z \; . \qquad (4.17)$$

Diese Methode kann auch für den Fall angewendet werden, in dem die Verteilungsfunktion zeitabhängig ist, d. h. der Verlauf der Verteilung ist vom Auswürfelungszeitpunkt abhängig. Die zeitabhängigen Verteilungsfunktionen treten bei der Simulation verallgemeinerter Poisson-Ankunftsprozesse auf. Dieser Sachverhalt wird im nächsten Unterkapitel bei der Beschreibung der Vorwärts-Integral-Methode näher erläutert.

Die Organisation eines Simulationslaufs wird mit Hilfe eines Rahmenprogramms festgelegt. Hierbei wird die gesamte Simulationszeitspanne aufgrund der statistischen Aussagesicherheit in eine Vorlaufphase und n Teiltests (Bild 4.6) unterteilt. Nach der Vorlaufphase befindet sich das System in dem eingeschwungenen Zustand. In diesem Zustand werden statistische Messungen zur Bestimmung charakteristischer Prozeß- bzw. Systemgrößen durchgeführt (z.B. Wartezeiten, Blockierungswahrscheinlichkeiten ...). Die Auswertung

von Meßdaten wird am Ende jedes Teiltests vorgenommen. Aus den Teiltestergebnissen, d. h. den n gewonnenen Stichproben, werden die charakteristischen Größen bestimmt. Mit den Stichproben-Meßwerten können auch Vertrauensintervalle als Maß für die statistische Aussagesicherheit berechnet werden.

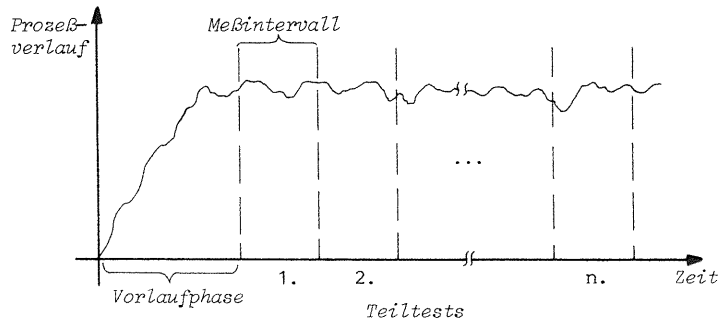


Bild 4.6 Meßmethode bei stationärer Simulation .

Bei der Simulation instationärer Vorgänge wird die beschriebene Meßmethode in einer modifizierten Form angewendet, da dort u. a. der Einschwingvorgang (z. B. die Vorlaufphase) untersucht werden soll. Ausgehend von einem Startvektor der Zustandswahrscheinlichkeiten des Systems zum Zeitpunkt $t_0 = 0$ wird instationär bis zu einem Zeitpunkt t_1 simuliert (Bild 4.7). Dieser Vorgang wird als ein Elementartest bezeichnet. Ähnlich wie im stationären Fall besteht ein Simulationslauf aus einer Anzahl von Teiltests. Jeder Teiltest umfaßt eine ausreichend große Anzahl von Elementartests.

Da die Systemcharakteristiken zeitabhängige Größen sind, werden sie in jedem Elementartest an einer vorgegebenen Reihe von Meßzeitpunkten gemessen. Die Meßzeitpunkte können beliebig angeordnet sein. Aus der Menge aller Meßwerte zu einem Meßzeitpunkt werden , analog zum stationären Fall, Ergebnisse und zugehörige Vertrauensintervalle gewonnen.

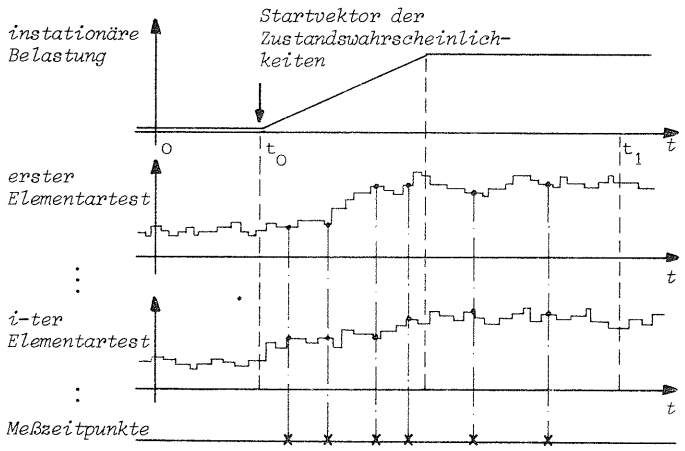


Bild 4.7 Meßmethode bei instationärer Simulation.

4.3.2 Simulation verallgemeinerter Poisson-Prozesse

Die Verkehrsströme, welche zusammen mit der Systemreaktion zu Überlastsituationen führen, stellen in der Regel instationäre Ankunftsprozesse dar. Die hier beschriebene Simulationstechnik beschränkt sich auf verallgemeinerte, instationäre Poisson-Prozesse, d. h. Poisson-Prozesse mit zeitlich veränderlicher, beliebiger Rate $\lambda(t)$.

a) Problemstellung

Der am häufigsten in der Literatur [34, 36] behandelte instationäre Poisson-Prozeß ist der Lastsprung, dessen Rate $\lambda(t)$ in Bild 4.8 dargestellt wird.

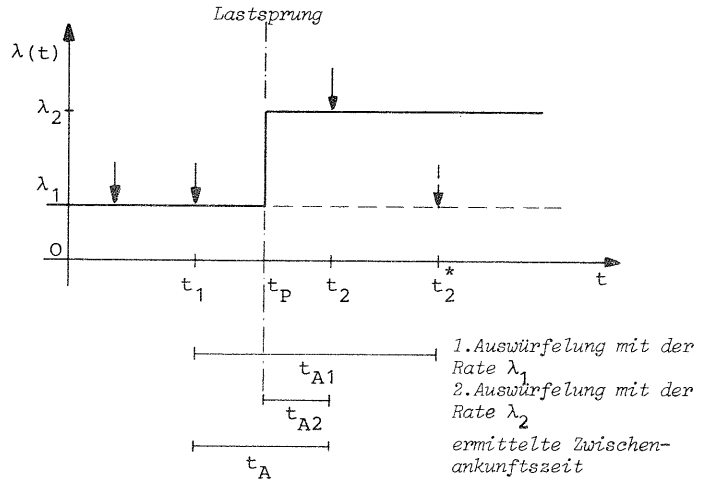


Bild 4.8 Zur Simulation eines Poisson-Lastsprungs.

Die Simulation eines Lastsprungs wird in Bild 4.8 gezeigt. Das hier beschriebene Verfahren beruht auf der Eigenschaft der Gedächtnislosigkeit von negativ exponentiell verteilten Zwischenankunftsabständen des Poisson-Prozesses (vgl. Kap. 5.1). Zum Zeitpunkt t_1 des letzten Ereignisses wird mit der Rate λ_1 das nächste Ereignis ausgewürfelt. Tritt das ermittelte Ereignis (zum Zeitpunkt t_2^*) erst nach dem Lastsprung ein, so wird es für ungültig angesehen, da sich inzwischen die Poisson-Rate geändert hat (von λ_1 auf λ_2). Aufgrund der gedächtnislosen Eigenschaft der negativ exponentiell verteilten Zeitintervalle zwischen den Ereignissen kann die Auswürfelung zum Zeitpunkt des Lastsprungs (t_P) nun mit der Rate λ_2 wiederholt werden. Das nächste Ereignis wird zum Zeitpunkt t_2 eintreten, und der tatsächliche Zwischenankunftsabstand über den Lastsprung ist t_A.

Nimmt die zeitabhängige Poisson-Rate $\lambda(t)$ einen beliebigen Verlauf an, so kann das beschriebene Verfahren nur approximativ

angewendet werden, indem $\lambda(t)$ durch einen stufenförmigen Kurvenverlauf angenähert wird. Diese Methode erfordert eine hinreichend genaue Stufen-Diskretisierung, die zur Folge hat, daß die Auswürfelung zur Bestimmung des nächsten Ereignisses sehr oft wiederholt werden muß. Die Simulation wird dadurch rechenzeitintensiv.

Zur Lösung dieses Problems wird nachfolgend beschriebene Vorwärts-Integral-Methode entwickelt.

b) Die Vorwärts-Integral-Methode zur Simulation verallgemeinerter Poisson-Prozesse

Bild 4.9 zeigt einen verallgemeinerten Poisson-Prozeß mit der zeitabhängigen Rate $\lambda(t)$. Ausgehend von einem Ereignis zum Zeitpunkt t_0 soll das nächste Ereignis simulativ bestimmt werden.

Die Zeitspanne bis zu einem betrachteten Zeitpunkt t wird nun in n Abschnitte der Länge Δt unterteilt, wobei zunächst die Rate während eines Zeitabschnittes als konstant angenommen wird.
Der i-te Abschnitt hat somit die Rate

$$\lambda_i = \lambda\left[t_0 + (i-1)\Delta t\right] \quad ; \; i=1,2,\ldots,n \; . \tag{4.18a}$$

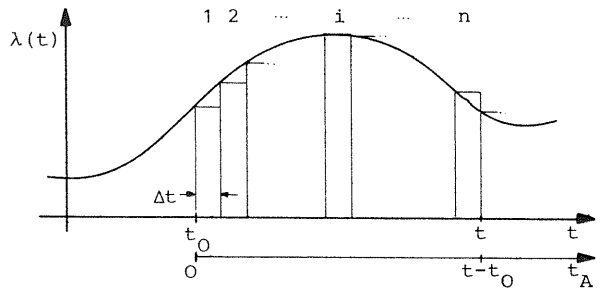


Bild 4.9 *Zur Herleitung der Vorwärts-Integral-Methode für die Simulation verallgemeinerter Poisson-Prozesse.*

Die Wahrscheinlichkeit, daß während des i-ten Zeitabschnittes kein Ereignis eintrifft, lautet:

$$P_0^{(i)} = e^{-\lambda_i \Delta t} \; . \tag{4.18b}$$

Aus Gl.(4.18) kann die Wahrscheinlichkeit dafür berechnet werden, daß kein Ereignis im Zeitintervall (t_0,t) eintrifft. Dies ist die Verbundwahrscheinlichkeit aller einzelner Wahrscheinlichkeiten $P_0^{(i)}$, die statistisch voneinander unabhängig sind. T_{A,t_0} sei die Zufallsvariable für die Zwischenankunftszeiten zwischen dem bekannten Ereignis zum Zeitpunkt t_0 und dem nächsten zu ermittelnden Ereignis. Es gilt:

$$P\{T_{A,t_0} > t_A\} = \prod_{i=1}^{n} P_0^{(i)} = \prod_{i=1}^{n} e^{-\lambda_i \Delta t}$$

$$= e^{-\Delta t \sum_{i=1}^{n} \lambda_i} \; . \tag{4.19}$$

Durch den Grenzübergang $n\to\infty, \Delta t\to 0$ geht die Summe im Exponenten von Gl.(4.19) in ein Integral über. Man erhält somit die komplementäre Verteilungsfunktion der Zwischenankunftszeiten zum Zeitpunkt t_0:

$$F^c_{t_0}(t_A) = P\{T_{A,t_0} > t_A\}$$

$$= \lim_{\substack{n\to\infty \\ \Delta t\to 0}} e^{-\Delta t \sum_{i=1}^{n} \lambda_i} = e^{-\int_{t_0}^{t_0+t_A} \lambda(t).dt} \; . \tag{4.20a}$$

Die zeitabhängige Zwischenankunftsverteilungsfunktion zum Zeitpunkt t_0 errechnet sich zu:

$$F_{t_0}(t_A) = P\{T_{A,t_0} \le t_A\} = 1 - e^{-\int_{t_0}^{t_0+t_A} \lambda(t).dt} \; . \tag{4.20b}$$

Für die Überlast- und Überlastabwehr-Untersuchungen werden häufig dynamische Systemantworten auf kurzzeitige Überlastung ermittelt. Dafür wird die instationäre Belastung mit impulsförmigen Überlastmustern modelliert, welche verallgemeinerte Poisson-Prozesse darstellen.

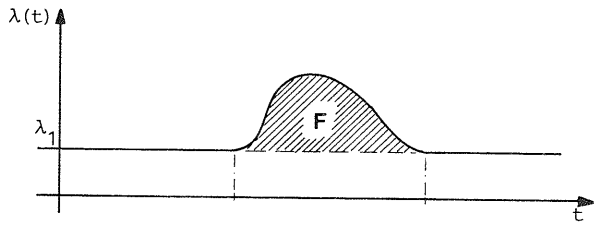


Bild 4.10 *Modellbildung kurzzeitiger Überlast mittels verallgemeinerter Poisson-Prozesse.*

F : *Überlast-Verkehrsvolumen*
λ_1 : *stationäre Grundlast*

Die zeitabhängige Poisson-Rate in Bild 4.10 besteht aus zwei Anteilen :

$$\lambda(t) = \lambda_1 + \lambda_{\ddot{u}}(t) \quad , \tag{4.21}$$

wobei $\lambda_{\ddot{u}}(t)$ den der Grundlast überlagerten Überlastverkehr darstellt.

Für derartige Überlastimpulse lautet die zeitabhängige Zwischenankunftsverteilungsfunktion aus Gl.(4.20b) und Gl.(4.21) :

$$F_{t_O}(t_A) = 1 - e^{-\lambda_1 t_A - K(t_O,t_A)} \quad , \tag{4.22}$$

mit

$$K(t_O,t_A) = \int_{t_O}^{t_O+t_A} \lambda_{\ddot{u}}(t).dt \; .$$

Der Term $K(t_O,t_A)$ wird durch die instationäre Überlastung $\lambda_{\ddot{u}}(t)$ erzeugt und wird im weiteren als Korrekturfunktion bezeichnet. Wird der Prozeß wieder stationär, so wird die Funktion $K(t_O,t_A)$ den Wert Null annehmen.

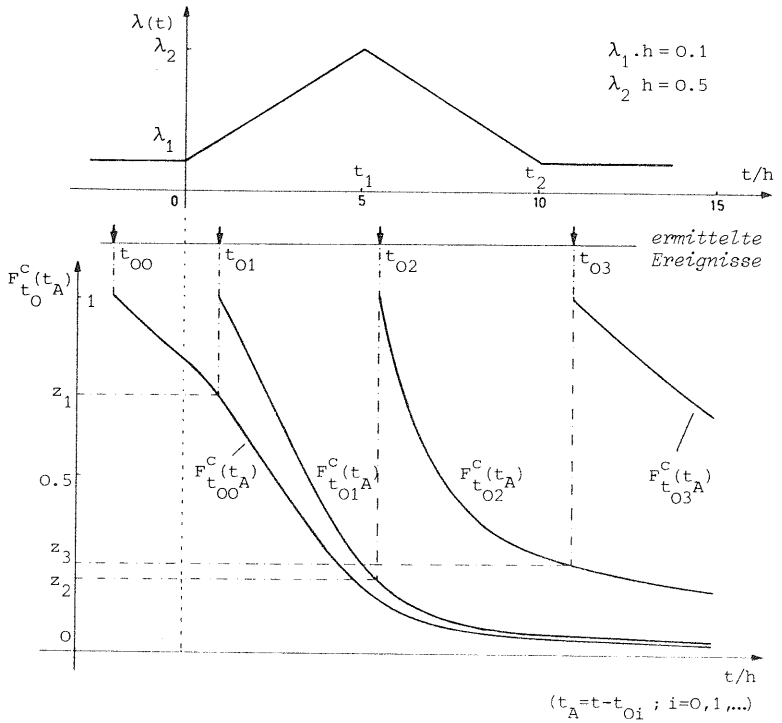


Bild 4.11 *Beispiel für die Vorwärts-Integral-Methode zur Simulation verallgemeinerter Poisson-Prozesse.*
(h : Zeiteinheit)

Im folgenden wird die Implementierung der Vorwärts-Integral-Methode gemäß Gl.(4.20) und (4.22) am Beispiel der dreieckförmigen Überlastung gezeigt. Die zeitabhängige Poisson-Rate des Dreieckimpulses in Bild 4.11 lautet :

$$\lambda(t) = \lambda_1 + \lambda_{\ddot{U}}(t) \ ,$$

mit

$$\lambda_{\ddot{U}}(t) = \begin{cases} \frac{t}{t_1}(\lambda_2-\lambda_1) & , \ 0 \le t < t_1 \ , \\ \frac{t_2-t}{t_2-t_1}(\lambda_2-\lambda_1) & , \ t_1 \le t < t_2 \ , \\ 0 & , \ \text{sonst} \ . \end{cases} \qquad (4.23)$$

Die zeitabhängige Zwischenankunfts-Verteilungsfunktion errechnet sich aus Gl.(4.22) zu :

$$F_{t_0}(t_A) = 1 - e^{-\lambda_1 t_A - K(t_0,t_A)} \ .$$

Dabei lautet die Korrekturfunktion für den Dreieckimpuls :

$\underline{t_0 < 0}$

$$K(t_0,t_A) = \begin{cases} 0 & , \ t_0 \le t < 0 \ , \\ \frac{\lambda_2-\lambda_1}{2t_1} t^2 & , \ 0 \le t < t_1 , \\ \frac{\lambda_2-\lambda_1}{2}\left(t_2 - \frac{(t_2-t)^2}{t_2-t_1}\right) & , \ t_1 \le t < t_2 , \\ \frac{\lambda_2-\lambda_1}{2} t_2 & , \ t \ge t_2 \ , \end{cases}$$

$\underline{0 \le t_0 < t_1}$

$$K(t_0,t_A) = \begin{cases} \frac{\lambda_2-\lambda_1}{2}\frac{t^2-t_0^2}{t_1} & , \ t_0 \le t < t_1 \ , \\ \frac{\lambda_2-\lambda_1}{2}\left(t_2 - \frac{t_0^2}{t_1} - \frac{(t_2-t)^2}{t_2-t_1}\right) & , \ t_1 \le t < t_2 \ , \\ \frac{\lambda_2-\lambda_1}{2}\left(t_2 - \frac{t_0^2}{t_1}\right) & , \ t \ge t_2 \ , \end{cases}$$

$\underline{t_1 \le t_0 < t_2}$

$$K(t_0,t_A) = \begin{cases} \frac{\lambda_2-\lambda_1}{2}\frac{(t_2-t_0)^2-(t_2-t)^2}{(t_2-t_1)} & , \ t_0 \le t < t_2 \ , \\ \frac{\lambda_2-\lambda_1}{2}\frac{(t_2-t_0)^2}{(t_2-t_1)} & , \ t \ge t_2 \ , \end{cases}$$

$\underline{t_0 \ge t_2}$

$$K(t_0,t_A) = 0 \ . \qquad \ldots(4.24)$$

In Bild 4.11 wird die Simulation der aufeinanderfolgenden Ereignisse zu den Zeitpunkten t_{01}, t_{02} und t_{03} gezeigt. Ausgehend vom letzten Ereignis zum Zeitpunkt t_{00} werden anhand der ausgewürfelten Zufallszahlen z_1, z_2 und z_3 sowie der Funktionen $F^c_{t_{0i}}(t_A)$ $(i = 0,1,2)$ die Ereigniszeitpunkte t_{01}, t_{02}, t_{03} nacheinander bestimmt. Dabei gilt jeweils $t_A = t - t_{0i}$ $(i=0,1,2)$.

Im Vergleich zu dem häufig benutzten Verfahren zur Simulation instationärer Ankunftsprozesse, bei dem die zeitabhängige Rate $\lambda(t)$ durch eine Treppenkurve - d.h. stückweise konstante Ankunftsraten - approximiert wird, weist die Vorwärts-Integral-Methode eine Reihe von Vorteilen in methodischer und implementierungstechnischer Hinsicht auf. Die Bestimmung von Ereignispunkten nach der Vorwärts-Integral-Methode ist im Gegensatz zu der Treppenkurve-Approximation exakt und unabhängig von den Diskretisierungsintervallen. Darüberhinaus erreicht man eine Reduzierung der Rechenzeit, da für ein Ereignis nur eine Zufallszahl ausgewürfelt wer-

den muß. Bei der Treppenkurve-Approximation, bedingt durch die stückweise konstanten Ankunftsraten und den in Bild 4.8 beschriebenen Abfragemechanismus, müssen für ein Ereignis häufig mehrere Zufallszahlen erzeugt werden.

5. MODELLE ZUR BESCHREIBUNG VON ÜBERLAST

In Fernsprechvermittlungssystemen besteht eine starke gegenseitige Beeinflussung zwischen der dynamischen Entwicklung der Verkehrsströme und der Leistungsfähigkeit eines Systems. Dieser Sachverhalt erfordert eine komplexe, rückkopplungsbehaftete Modellbildung zur Erfassung stationärer und dynamischer stochastischer Systemreaktionen sowie Regelmechanismen, mit denen die Entwicklung von Überlastsituationen erklärt werden kann. Anhand einiger entwickelter Grundmodelle werden in diesem Kapitel die wesentlichen Interaktionen zwischen System- und Teilnehmerverhalten mit dem Ziel untersucht:

- Verkehrsströme in Vermittlungssystemen statistisch zu beschreiben
- das Zustandekommen von Überlastsituationen qualitativ und quantitativ zu erfassen.

Die Modelle werden klassifiziert in:

- Modelle und Modellelemente für Verkehrsströme:
  Die Systemrückwirkung wird hier nicht explizit berücksichtigt. Dazu zählen exakte und approximative Beschreibungsmethoden für Ruf- und Teilrufverkehre, die für stationäre und instationäre Modelluntersuchungen angewendet werden.
- Modelle zur Beschreibung der Überlastsituationen:
  Mit dieser Klasse von Modellen werden dynamische Entwicklungen von Überlastsituationen untersucht, wobei das Teilnehmerverhalten und die Systemrückwirkung auf Eingangsverkehrsströme berücksichtigt werden.

5.1 Rückwirkungsfreie Modelle zur Beschreibung von Überlastsituationen

Zur Charakterisierung der Überlastverkehrsströme sind stationäre und instationäre Ankunftsprozesse erforderlich, die Ruf-, Teilprozeß- und Teilrufverkehrsströme beschreiben. Ein Ankunftsprozeß beschreibt die zeitliche Reihenfolge bzw. die Eintreffzeitpunkte von Anforderungen am Eingang eines Verkehrsmodells bzw. eines Systems.

Abhängig vom betrachteten System (z.B. dem Prozessor, den Software-Verarbeitungseinheiten in einer Software-Maschine oder dem gesamten Vermittlungssystem) wird die Modellbildung des Ankunftsprozesses auf der Ruf-, Teilprozeß- oder Teilrufebene durchgeführt.

Für die analytische Behandlung dieser Prozesse werden einige Grundbeziehungen der Prozeßtheorie benötigt [5,6,11], die nachfolgend erörtert werden.

5.1.1 Allgemeines über die Theorie der Erneuerungsprozesse

Ein Ankunftsprozeß heißt Erneuerungsprozeß (Rekurrenter Prozeß), wenn die Zeitintervalle zwischen aufeinanderfolgenden Ereignissen unabhängig voneinander und identisch verteilt sind. Die Ankunftsabstände mit der Zufallsvariable T_A werden durch eine Verteilungsfunktion beschrieben

$$F_A(t) = P\{T_A \leq t\}. \qquad (5.1a)$$

Die zugehörige Verteilungsdichtefunktion lautet

$$f_A(t) = \frac{d}{dt} F_A(t). \qquad (5.1b)$$

Für eine einfache analytische Handhabung von Verteilungsfunktionen wird die Laplace-Stieltjes-Transformation (LST) eingeführt, welche nachfolgend im Zusammenhang mit der bekannten Laplace-Transformation definiert wird

$$\Phi_A(s) = \int_{0^-}^{\infty} e^{-st}\, dF_A(t) = \int_{0^-}^{\infty} e^{-st} f_A(t).dt. \qquad (5.2a)$$

$\Phi_A(s)$ ist die Laplace-Stieltjes-Transformierte von $F_A(t)$ und ist identisch mit der Laplace-Transformierten von $f_A(t)$:

$$F_A(t) \circ\!\!-\!\!\!-\!\!\!\xrightarrow{LST}\!\!\!-\!\!\bullet\ \Phi_A(s) \qquad \Phi_A(s) = LST\{F_A(t)\}$$

$$\text{oder} \qquad \ldots(5.2b)$$

$$f_A(t) \circ\!\!-\!\!\!-\!\!\!\xrightarrow{LT}\!\!\!-\!\!\bullet\ \Phi_A(s) \qquad \Phi_A(s) = LT\{f_A(t)\}.$$

Betrachtet man nun den Prozeß zu einem beliebigen Zeitpunkt (Bild 5.1), so entstehen zwei neue Zufallsvariablen, die für die Beschreibung von Erneuerungsprozessen von Bedeutung sind:

- Die Vorwärts-Rekurrenzzeit T_A^v: zufälliges Zeitintervall vom Betrachtungszeitpunkt bis zum nächsten Ereignis.
- Die Rückwärts-Rekurrenzzeit T_A^r: zufälliges Zeitintervall vom letzten Ereignis bis zum Betrachtungszeitpunkt.

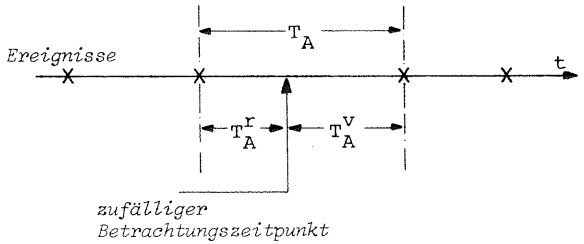


Bild 5.1 *Rekurrenzzeiten eines Ankunftsprozesses.*
T_A: *Zwischenankunftszeit*
T_A^r: *Rückwärts-Rekurrenzzeit*
T_A^v: *Vorwärts-Rekurrenzzeit*

Die Erneuerungstheorie [5,6] liefert folgende Beziehung zwischen Verteilungsfunktionen der Vorwärts-Rekurrenzzeit T_A^v und der Zwischenankunftszeit T_A

$$f_A^v(t) = \frac{1 - F_A(t)}{E[T_A]} = \lambda(1 - F_A(t)), \qquad (5.3)$$

wobei

$f_A^V(t)$ Verteilungsdichtefunktion der Vorwärts-Rekurrenzzeit

$F_A(t)$ Verteilungsfunktion der Zwischenankunftszeiten

$E[T_A]$ Mittlerer Ankunftsabstand

$\lambda = \frac{1}{E[T_A]}$ Mittlere Ankunftsrate des Prozesses.

Die Gl. (5.3) lautet nach der Laplace-Stieltjes-Transformation:

$$\Phi_A^V(s) = \frac{\lambda}{s}(1 - \Phi_A(s)) \quad , \tag{5.4}$$

mit $\Phi_A^V(s) = LT\{f_A^V(t)\}$

und $\Phi_A(s) = LST\{F_A(t)\}$.

Im Falle einer negativ exponentiellen Verteilung ist die Verteilung der Vorwärts-Rekurrenzzeit T_A^V identisch mit der Verteilung von T_A. Dies ist aus der Markoff-Eigenschaft der negativ exponentiellen Verteilung zu erwarten. Aus Gl. (5.4) erhält man für die negativ exponentielle Verteilung:

$$\Phi_A^V(s) = \frac{\lambda}{s}(1 - \Phi_A(s)) = \frac{\lambda}{s}(1 - LST\{1 - e^{-\lambda t}\})$$

$$= \frac{\lambda}{s}(1 - \frac{\lambda}{s+\lambda}) = \frac{\lambda}{s+\lambda} = \Phi_A(s). \tag{5.5}$$

5.1.2 Modellelemente für Ruf-, Teilprozeß- und Teilrufverkehrsströme

Für die verschiedenen Modellebenen - Ruf-, Teilprozeß- und Teilrufebene - werden z.T. unterschiedliche Prozeßbeschreibungen benötigt. Sie stellen Modellelemente dar, welche für die in dieser Arbeit behandelten Modellansätze Eingangsprozesse bilden.

a) Rufverkehr

In den meisten Modellbildungen wird für die Beschreibung des Rufverkehrs der stationäre Poisson-Prozeß angewendet. Diese Annahme, die insbesondere bei einer großen Anzahl von angeschlossenen Teilnehmern eine gute Näherung darstellt, wird auch durch Messungen an Fernsprechvermittlungssystemen in Hauptverkehrsstunden bestätigt.

Der stationäre Poisson-Prozeß erzeugt Rufe mit negativ exponentiell verteilten Ankunftsabständen T_A

$$F_A(t) = P\{T_A \leq t\} = 1 - e^{-\lambda t} \quad , \tag{5.6}$$

mit T_A: Zufallsvariable für den Ankunftsabstand

$\lambda = 1/E[T_A]$: Ankunftsrate.

Die Wahrscheinlichkeit, daß n Rufe in einem Zeitintervall T eintreffen, lautet:

$$g_n(T) = \frac{(\lambda T)^n}{n!} e^{-\lambda T}. \tag{5.7}$$

Zur Beschreibung des Rufverkehrs in Überlastsituationen sind instationäre Ankunftsprozesse erforderlich, die näherungsweise die zeitabhängigen Rufverkehrsströme modellieren. Dabei werden in erster Näherung die Systemrückwirkung auf den Prozeß sowie mögliche Abhängigkeiten zwischen den Rufen vernachlässigt.

Eine Möglichkeit zur Beschreibung instationärer Verkehrsströme bietet der verallgemeinerte Poisson-Prozeß, der durch eine zeitabhängige Ankunftsrate $\lambda(t)$ charakterisiert wird (vgl. Gl. 4.10). In dieser Arbeit wird der verallgemeinerte Poisson-Prozeß simulativ und analytisch behandelt. Kapitel 4.3 stellt die Vorwärts-Integral-Methode zur zeittreuen Simulation verallgemeinerter Poisson-Prozesse vor. Im Modellansatz von Kapitel 5.2.2 wird der Prozeß zur analytischen Behandlung von impulsförmigen Überlast-Verkehrsmustern angewendet.

b) Teilprozeß-Verkehr

Anforderungen zur Aktivierung von Teilprozessen, die für die Bearbeitung und Überwachung aktiver Rufe im System erforderlich sind, bilden den Teilprozeß-Verkehr, der z.B. zur Modellbildung von Software-Maschinen und rufbezogenen Datenblöcken benötigt wird. Da der Aktivierungszeitpunkt eines Teilprozesses häufig von der echtzeitmäßigen Verarbeitung vorhergehender Teilprozesse abhängig ist und die Teilprozesse eines Rufes von unterschiedlicher Dauer sind, ist es nicht einfach, eine geeignete Prozeßbeschreibung des Teilprozeß-Verkehrs zu definieren. Folgende Aspekte werden bei der Modellbildung berücksichtigt:

- Ist die Anzahl aktiver Rufe im System genügend groß, so kann der Teilprozeß-Verkehr mit einem Erneuerungsprozeß, z.B. dem Poisson-Prozeß, modelliert werden.
- Falls die Modelluntersuchung eine detaillierte Betrachtung der Teilprozeß-Aktivierung und -Verarbeitung erfordert, muß die gegenseitige Beeinflussung von Teilprozessen berücksichtigt werden. Das Modell in Kap. 6.2.1 stellt ein Beispiel dar, bei dem die serielle und parallele Teilprozeß-Aktivierung in Software-Maschinen untersucht wird.

c) Teilruf-Verkehr

Die Definition von Teilrufen hängt davon ab, welche vermittlungstechnische Einrichtung man zur jeweiligen Modellbildung heranzieht. Wird der zentrale oder dezentrale Vermittlungsrechner

betrachtet, so sind Teilrufe die zur Verarbeitung eines Rufes erforderlichen Steuerungsaufrufe. Modelliert man die periphere Schnittstelle mit den für die Ein- und Ausgabe bestimmten vermittlungstechnischen Signalen, so werden diese Signale als Teilrufe betrachtet (vgl. Kap. 6).

Der Teilrufverkehr hat eine von der Anzahl aktiver Rufe im System abhängige Intensität. Überdies sind die Zwischenankunftsabstände der Teilrufe, die zu einem aktiven Ruf gehören, voneinander abhängig, so daß der Teilrufstrom mit einem Erneuerungsprozeß nur näherungsweise zu beschreiben ist. Die Unabhängigkeitsannahme für den Teilrufstrom ist nur dann erlaubt, wenn die mittlere Anzahl der aktiven Rufe im System genügend groß ist.

Abhängig von der Modellierungstiefe können folgende Elemente zur Modellbildung herangezogen werden:

(1) Ereignisketten

Ein aktiver Ruf, der eine Folge von Ereignissen bzw. Teilrufen im System erzeugt, kann mit einem Echtzeit-Prozeß beschrieben werden. Der Prozeß besteht aus einer endlichen Anzahl n der Zustände. Die erzeugten Ereignisse führen zu Zustandsänderungen, die in Bild 5.2 mit Hilfe der Beschreibungssprache SDL dargestellt werden (SDL: functional specification and description language [13]).

Wird ein Ruf aktiviert, so befindet sich der Prozeß im Zustand "frei", der z.B. mit Null (0) numeriert wird. Mit den Verzweigungswahrscheinlichkeiten p_{0i}, $i = 0,\ldots,n$, können alle Ruftypen (interne oder externe, ankommende oder abgehende Verbindungen,...) charakterisiert werden. Das erste Ereignis "Belegung" (Ereignis 0,i; i = 0,..,n) initiiert die Ereigniskette. Mit den vorausgesetzten oder gemessenen Verzweigungswahrscheinlichkeiten $p_{i,j}$ sowie den Verweilzeiten $T_{i,j}$ $(i,j = 0,1,\ldots,n)$ können Ereignisketten modelliert werden, wobei sowohl System- als auch Teilnehmerverhalten berücksichtigt werden kann [34, 48]. Die Zufallsvariablen $T_{i,j}$ stellen dabei die teilruf- und zustandsabhängigen Zwischenankunftsintervalle der Teilrufe dar.

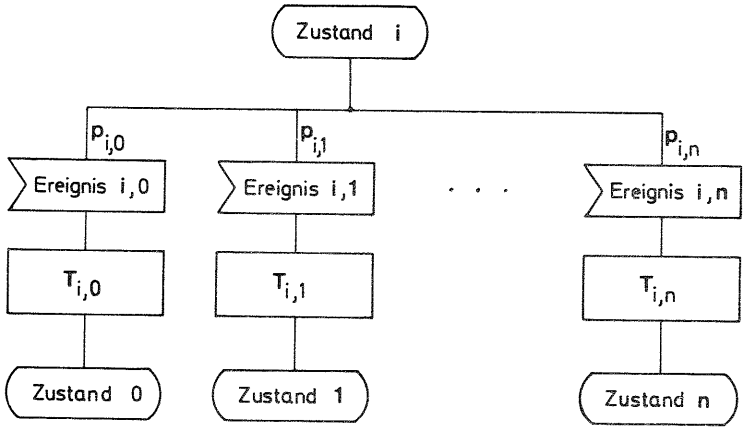


Bild 5.2 *Erzeugungsmodell von Teilrufen für die Ereignisketten-Beschreibung.*

Eine Ereigniskette geht zu Ende, wenn der Zustand "frei" wieder erreicht wird. Die typischen Ereignisketten, z.B. bezüglich des Teilnehmerverhaltens

- Abheben → Auflegen
- Abheben → Ziffern → Auflegen
- Abheben → Ziffern → Auflegen nach Rufton
- Abheben → Ziffern → B melden → Auflegen nach Gespräch

...,

können durch entsprechende Festlegung der Verzweigungswahrscheinlichkeiten, der Werte für den Wähltonverzug sowie der Zeitintervalle zwischen Ziffern usw. erzeugt werden.

(2) Verzweigte Prozesse (branching processes)

Der Ansatz für diese Beschreibung beruht auf der Beobachtung in rechnergesteuerten Vermittlungssystemen, daß die Anzahl der erzeugten Teilrufe pro Ruf sehr groß ist. Als Folge wird die Teilruferzeugung pro Ruf (z.B. in der Rufaufbauphase) als ein Erneuerungsprozeß angenommen. Die Überlagerung aller im aktiven Zustand befindlichen Teilruf-Erzeugungsprozesse ergibt den Teilrufverkehr. Der Rufprozeß "verzweigt" sich demgemäß zu Teilruf-Erzeugungsprozessen.
Eine detaillierte Darstellung dieses Modellbildungsansatzes findet sich in Kap. 6.2.2, in dem der Teilruf-Erzeugungsprozeß (bezogen auf die Aufbauphase eines Rufes) als Poisson-Prozeß angenommen wird. Es handelt sich hierbei um eine modifizierte Form des verzweigten Poisson-Prozesses (branching Poisson-process).

(3) Erneuerungsprozesse

Ist die Intensität des Teilrufverkehrs genügend groß, so verringert sich die gegenseitige Abhängigkeit zwischen den Zwischenankunftsabständen der Teilrufe. Dies erlaubt eine Beschreibung des Teilrufverkehrs mit Erneuerungsprozessen. Im nächsten Unterkapitel wird der Teilrufstrom durch einen geschalteten Poisson-Prozeß approximiert, wobei eine Erneuerungsannahme gemacht wird.

5.1.3 Erneuerungsapproximation für den geschalteten Poisson-Prozeß

a) Näherungsweise Beschreibung des Teilrufverkehrs durch den geschalteten Poisson-Prozeß

In Vermittlungssystemen mit modularer Struktur ist die Anzahl der Teilrufe bzw. Steuerungsaufrufe pro Ruf im Vergleich zu konventionellen Systemen relativ groß. Überdies ist die Intensität des Teilrufverkehrs stark abhängig von der Anzahl aktiver Teilprozesse bzw. Rufe im System, so daß der Teilrufprozeß mit stationären Erneuerungsprozessen nur näherungsweise zu beschreiben ist. In Untersuchungen zur Leistungsbeurteilung derartiger Systeme erfordert der Teilrufstrom eine möglichst exakte statistische Darstellung. Diesbezüglich werden in der vorliegenden Arbeit zwei Möglichkeiten vorgestellt und untersucht:

- Beschreibung des Teilrufverkehrs mit Hilfe einer wirklichkeitsnahen Modellbildung, wobei Ruf- und Teilrufprozesse einschließlich der Teilruferzeugung zusammen betrachtet werden. Dieser Modellansatz wird in Kap. 6.2.2 behandelt.
- Approximative Beschreibung mit einem Erneuerungsprozeß. Da der Teilrufverkehr hohe Varianz und überdies Abhängigkeit zwischen den Prozeßabschnitten aufweist, soll der approximierende Prozeß ebenfalls diese Eigenschaften besitzen. In diesem Zusammenhang wird nachfolgend die Approximation des Teilrufverkehrs mit Hilfe des geschalteten Poisson-Prozesses untersucht.

Für den geschalteten Poisson-Prozeß wird eine Verteilungsfunktion der Zwischenankunftsabstände entwickelt, wobei eine Approximation mittels der Erneuerungsannahme vorgenommen wird. Die gewonnene Verteilung weist eine hyperexponentielle Charakteristik auf. Dies entspricht dem Verlauf der in [48] simulativ ermittelten Verteilungsfunktion für den Teilrufstrom.

b) Definition des geschalteten Poisson-Prozesses

Der hier vorgestellte Modellansatz befaßt sich mit dem verallgemeinerten geschalteten Poisson-Prozeß (SPP: Switched Poisson Process). Dieser Prozeß entsteht durch ein zufallsmäßiges, alternierendes Umschalten zwischen zwei Poisson-Prozessen, die mit den Ankunftsraten λ_1 und λ_2 gekennzeichnet sind. (Bild 5.3). Die Aufenthaltszeiten T_1 und T_2 des Prozesses in den zwei Phasen sind dabei Zufallsvariablen (ZV) mit beliebigen Verteilungsfunktionen.

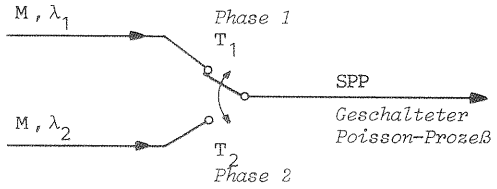


Bild 5.3 *Erzeugungsmodell des geschalteten Poisson-Prozesses.*

Es finden sich in der Literatur einige Modellansätze und Untersuchungen, in denen der geschaltete Poisson-Prozeß und seine modifizierten Formen angewendet werden. In [52-54] wird der gewöhnliche geschaltete Poisson-Prozeß diskutiert, der durch negativ exponentiell verteilte Aufenthaltszeiten charakterisiert wird. Infolgedessen wird dieser Prozeß auch Markoff-modulierter Poisson-Prozeß (MMP Markov-modulated Poisson process [52]) genannt, der in [49,51,53,54] als Eingangsprozeß von Grundmodellen der Verkehrstheorie untersucht wird.

Einen Sonderfall des geschalteten Poisson-Prozesses bildet der unterbrochene Poisson-Prozeß (IPP: Interrupted Poisson Process [50]). Dieser Prozeß entsteht aus dem SPP, bei dem eine Poisson-Rate (λ_1 oder λ_2) den Wert Null annimmt.

Durch die Festlegung der Parameter läßt sich der geschaltete Poisson-Prozeß zwischen dem gewöhnlichen Poisson-Prozeß und dem unterbrochenen Poisson-Prozeß variieren. Während die zwei genannten Grenzfälle die Erneuerungseigenschaft aufweisen, ist SPP kein Erneuerungsprozeß. Hier wird für den SPP eine Erneuerungsannahme vorgestellt, die anhand des Warte-Verlust-Systems SPP/M/1-S hinsichtlich der Approximationsgenauigkeit untersucht wird.

c) Parameterfestlegung

Folgende Notation wird bei der Darstellung der in diesem Unterkapitel auftretenden Zufallsvariablen benutzt:

T Zufallsvariable.

T^v Vorwärts-Rekurrenzzeit von T.

$F(t) = P\{T \leq t\}$
Verteilungsfunktion der Zufallsvariable T.

$f(t) = \frac{dF(t)}{dt}$
Verteilungsdichtefunktion der Zufallsvariable T.

$\Phi(s) = LT\{f(t)\} = LST\{F(t)\}$
Laplace-Stieltjes-Transformierte von F(t) bzw.
Laplace-Transformierte von f(t). ...(5.8)

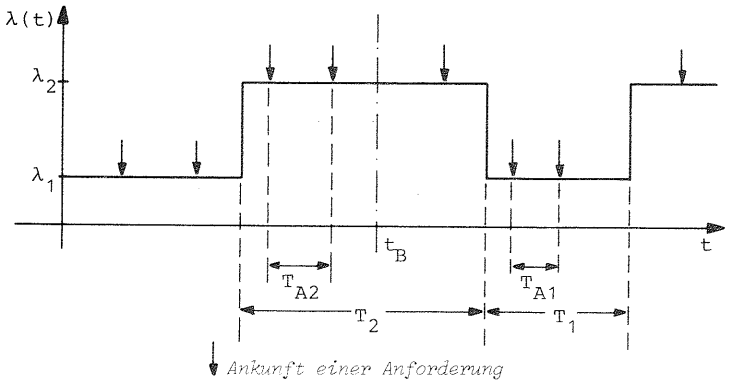


Bild 5.4 *Verlauf des geschalteten Poisson-Prozesses.*

Anhand des Prozeßverlaufs, der in Bild 5.4 gezeigt wird, werden die Parameter des geschalteten Poisson-Prozesses erläutert. Durch die folgenden Parameter wird der Prozeß vollständig gekennzeichnet:

T_1 ZV für die Aufenthaltsdauer in Phase 1
T_2 ZV für die Aufenthaltsdauer in Phase 2
T_{A1} ZV für den Zwischenankunftsabstand in Phase 1
T_{A2} ZV für den Zwischenankunftsabstand in Phase 2.

Aufgrund der Prozeßdefinition nach Bild 5.3 ergibt sich:

$$F_{A1}(t) = P\{T_{A1} \leq t\} = 1 - e^{-\lambda_1 t}; \quad E[T_{A1}] = \frac{1}{\lambda_1}, \tag{5.9a}$$

$$F_{A2}(t) = P\{T_{A2} \leq t\} = 1 - e^{-\lambda_2 t}; \quad E[T_{A2}] = \frac{1}{\lambda_2}. \tag{5.9b}$$

Die Aufenthaltszeiten besitzen die Mittelwerte:

$$E[T_1] = h_1 = \frac{1}{\omega_1}, \tag{5.10a}$$

$$E[T_2] = h_2 = \frac{1}{\omega_2}. \tag{5.10b}$$

Aus den vier Grundparametern des geschalteten Poisson-Prozesses lassen sich folgende Parameter herleiten:

- Mittlere Rate der Anforderungen

$$\lambda = \frac{\lambda_1 h_1 + \lambda_2 h_2}{h_1 + h_2} = \frac{\lambda_1 \omega_2 + \lambda_2 \omega_1}{\omega_2 + \omega_1}. \tag{5.11}$$

- Bezeichnet man eine Phase 1 und eine Phase 2 als eine Periode des geschalteten Poisson-Prozesses, so errechnet sich die mittlere Anzahl der Anforderungen pro Periode zu

$$n_O = \lambda_1 h_1 + \lambda_2 h_2 = \frac{\lambda_1}{\omega_1} + \frac{\lambda_2}{\omega_2} , \qquad (5.12)$$

n_O kennzeichnet zusammen mit λ die Schalthäufigkeit des Prozesses.

- Verhältnis der Phasendauer

$$\Theta = \frac{h_2}{h_1} = \frac{\omega_1}{\omega_2} . \qquad (5.13)$$

- Überlastfaktor

$$\gamma = \frac{\lambda_2}{\lambda} . \qquad (5.14)$$

Der Überlastfaktor variiert zwischen folgenden Grenzwerten, welche den Grenzprozessen entsprechen:

- Poisson-Prozeß :

$$\lambda_1 = \lambda_2 = \lambda \rightarrow \gamma_{min} = 1 .$$

- Unterbrochener Poisson-Prozeß (IPP) :

$$\lambda_1 = 0 \rightarrow \gamma_{max} = \frac{\omega_1+\omega_2}{\lambda_2 \omega_1} \lambda_2 = 1 + \frac{1}{\Theta}.$$

d) Approximative Prozeßbeschreibung mittels Erneuerungsannahme

Die Herleitung einer Verteilungsfunktion zur Approximation des geschalteten Poisson-Prozesses erfolgt in zwei Schritten:

1) Berechnung der Verteilungsfunktion der Vorwärts-Rekurrenzzeit (im Bildbereich der Laplace-Stieltjes-Transformation)
2) Berechnung der Verteilungsfunktion von Zwischenankunftszeiten der Ereignisse mit Hilfe der Erneuerungsannahme.

Zur Bestimmung der Vorwärts-Rekurrenzzeit-Verteilung wird der Prozeß zu einem beliebigen, zufällig herausgegriffenen Zeitpunkt (Beobachtungszeitpunkt t_B, s.Bild 5.4) betrachtet. Von diesem Zeitpunkt an bis zur nächsten eintreffenden Anforderung wird der Prozeßverlauf verfolgt. Die Zeitspanne zwischen dem Beobachtungszeitpunkt und der nächsten Anforderung stellt die Vorwärts-Rekurrenzzeit mit der ZV T^V dar. Da der Beobachtungszeitpunkt t_B zufällig gewählt wird, liegt t_B in einer Phase 1 mit der Wahrscheinlichkeit:

$$p_1 = \frac{h_1}{h_1 + h_2} = \frac{\frac{1}{\omega_1}}{\frac{1}{\omega_1} + \frac{1}{\omega_2}} = \frac{\omega_2}{\omega_1 + \omega_2} , \qquad (5.15a)$$

und analog für die Phase 2

$$p_2 = \frac{\omega_1}{\omega_1 + \omega_2}. \qquad (5.15b)$$

Der Betrachtungszeitpunkt t_B liege nun in einer Phase 1. Bild 5.5 illustriert exemplarisch drei Realisierungen für die Vorwärts-Rekurrenzzeit T^V. Von t_B an findet ein Wettlauf zwischen der Restphase T^V_{A1} (Vorwärts-Rekurrenzzeit von T_{A1}) und der Restphase T^V_1 statt, so daß die Zeit bis zur nächsten Zustandsänderung des Prozesses mit der ZV $T^V_{m1} = \min\{T^V_{A1}, T^V_1\}$ beschrieben werden kann.

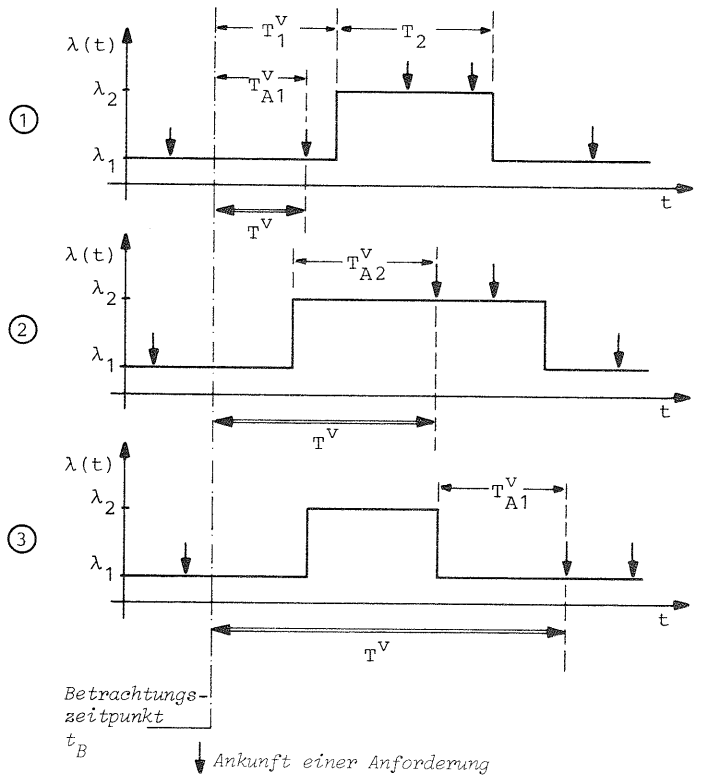


Bild 5.5 *Zur Herleitung der Vorwärts-Rekurrenzzeit-Verteilungsfunktion des geschalteten Poisson-Prozesses.*

Nach Ablauf der Zeit T_{m1}^{V} kann

- eine Anforderung der laufenden Phase 1 eintreffen (Fall ① in Bild 5.5). Die Vorwärts-Rekurrenzzeit beträgt in diesem Falle

  $$T^{V} = T_{m1}^{V}$$

- die laufende Phase 1 endigen, bevor eine Anforderung eintrifft. Der Prozeßverlauf wird weiter verfolgt.

Bis zur nächsten Zustandsänderung findet ein Wettlauf zwischen der Phase T_2 und der Restphase T_{A2}^{V} statt, so daß die Zeitspanne $T_{m2} = \min\{T_{A2}^{V}, T_2\}$ beträgt. Nach Ablauf dieser Zeit kann

- eine Anforderung der betrachteten Phase 2 eintreffen (Fall ② in Bild 5.5). Die Zeit T^{V} dauert insgesamt

  $$T^{V} = T_{m1}^{V} + T_{m2}$$

- die betrachtete Phase 2 endigen, ohne daß eine Anforderung in dieser Phase eintrifft. In diesem Fall wird eine neue Phase 1 angefangen.

Ähnlich findet nun ein Wettlauf zwischen einer Restphase T_{A1}^{V} und einer Phase T_1 statt, so daß eine Zeitspanne mit der ZV $T_{m1} = \{T_{A1}^{V}, T_1\}$ bis zur nächsten Zustandsänderung andauert. Am Ende dieser Zeitspanne kann

- eine Anforderung der betrachteten Phase 1 eintreffen (Fall ③ in Bild 5.5). Die Vorwärts-Rekurrenzzeit beträgt:

  $$T^{V} = T_{m1}^{V} + T_{m2} + T_{m1}$$

- die Phase 1 endigen. Die Beobachtung des Prozesses kann analog fortgesetzt werden.

Berücksichtigt man alle kombinatorischen Möglichkeiten, so kann die Vorwärts-Rekurrenzzeit T^{V} wie in Bild 5.6 schematisch dargestellt werden. Bedingung dabei ist, daß der Beobachtungszeitpunkt in einer Phase 1 liegt.

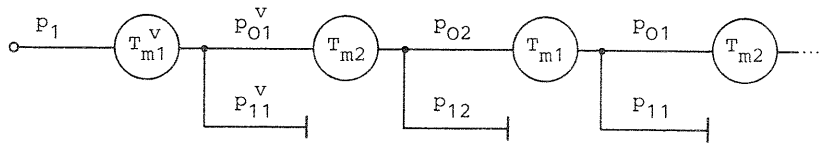


Bild 5.6 *Zur Betrachtung der Vorwärts-Rekurrenzzeit.*

zufälliger Betrachtungszeitpunkt
Ankunft der nächsten Anforderung

Die Bedeutung der in Bild 5.6 dargestellten Wahrscheinlichkeiten kann gemäß folgender Notation erklärt werden (i = 1,2):

p_i	Beobachtungszeitpunkt liegt in Phase i.
p_{Oi}^v	keine Anforderung trifft in der Restphase i ein.
$p_{1i}^v = 1 - p_{Oi}^v$	mindestens eine Anforderung trifft in der Restphase i ein.
p_{Oi}	keine Anforderung trifft in der Phase i ein.
$p_{1i} = 1 - p_{Oi}$	mindestens eine Anforderung trifft in der Phase i ein.

Betrachtet man nun den zufälligen Beobachtungszeitpunkt in beiden Phasen, so erhält man die Darstellung für T^v in Bild 5.7.

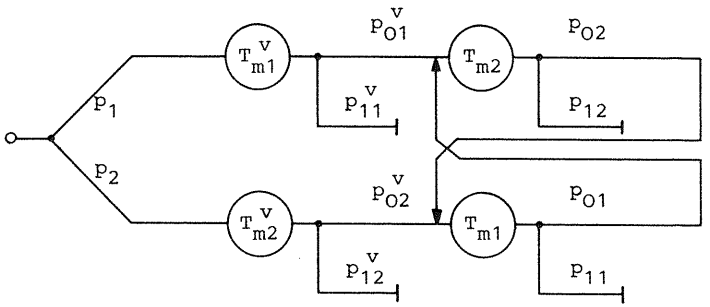


Bild 5.7 *Phasendarstellung der Vorwärts-Rekurrenzzeit für den geschalteten Poisson-Prozeß.*

o——— *zufälliger Betrachtungszeitpunkt*

———| *Ankunft der nächsten Anforderung*

Aus Bild 5.7 ist ersichtlich, daß sich die Laplace-Stieltjes-Transformierte $\Phi^v(s)$ der Vorwärts-Rekurrenzzeit-Verteilungsfunktion $F^v(t)$ aus zwei Anteilen zusammensetzt, die aus der Betrachtung des oberen und des unteren Zweiges herrühren.

$$\Phi^v(s) = \Phi_1^v(s) + \Phi_2^v(s) \quad . \tag{5.16}$$

Am Beispiel des oberen Zweiges wird nun die Berechnung von $\Phi_1^v(s)$ durchgeführt (Bild 5.8).

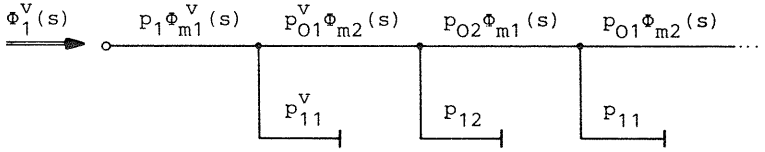


Bild 5.8 *Zur Berechnung der Vorwärts-Rekurrenzzeit-Verteilungsfunktion im Laplace-Stieltjes-Bereich.*

Gemäß der Darstellung in Bild 5.8 errechnet sich $\Phi_1^v(s)$ zu:

$$\Phi_1^v(s) = p_1\Phi_{m1}^v(s)\Big(p_{11}^v + p_{01}^v\Phi_{m2}(s) \cdot \Big[\underbrace{p_{12}}_{a} + \underbrace{p_{02}\Phi_{m1}(s)}_{b}\underbrace{(p_{11}}_{c} + \underbrace{p_{01}\Phi_{m2}(s)}_{d})\underbrace{(p_{12}}_{a} + \underbrace{p_{02}\Phi_{m1}(s)}_{b})\underbrace{(p_{11}}_{c} + \ldots)))\Big]\Big).$$

Mit

$$[\ldots] = a+bc+(bd)a+(bd)bc+(bd)^2a+(bd)^2bc+(bd)^3a+\ldots$$

$$= (a+bc)\sum_{i=0}^{\infty}(bd)^i = \frac{a+bc}{1-bd}$$

erhält man

$$\Phi_1^v(s) = p_1\Phi_{m1}^v(s)\left(p_{11}^v+p_{O1}^v\Phi_{m2}(s)\frac{p_{12}+p_{O2}p_{11}\Phi_{m1}(s)}{1-p_{O1}p_{O2}\Phi_{m1}(s)\Phi_{m2}(s)}\right). \qquad (5.17)$$

Der zweite Anteil $\Phi_2^v(s)$ wird analog aus der Betrachtung des unteren Zweiges in Bild 5.7 berechnet. Schließlich lautet die Laplace-Stieltjes-Transformierte der Vorwärts-Rekurrenzzeit-Verteilungsfunktion gemäß Gl. (5.16):

$$\Phi^v(s) = p_1\Phi_{m1}^v(s)\left(p_{11}^v+p_{O1}^v\Phi_{m2}(s)\frac{p_{12}+p_{O2}p_{11}\Phi_{m1}(s)}{1-p_{O1}p_{O2}\Phi_{m1}(s)\Phi_{m2}(s)}\right)$$

$$+p_2\Phi_{m2}^v(s)\left(p_{12}^v+p_{O2}^v\Phi_{m1}(s)\frac{p_{11}+p_{O1}p_{12}\Phi_{m2}(s)}{1-p_{O1}p_{O2}\Phi_{m1}(s)\Phi_{m2}(s)}\right). \qquad (5.18)$$

Mit Hilfe der Erneuerungsannahme (vgl.Kap. 5.1.1) kann nun die Laplace-Stieltjes-Transformierte der Verteilungsfunktion von Zwischenankunftszeiten angegeben werden:

$$\Phi(s) = 1 - \frac{s}{\lambda}\Phi^v(s). \qquad (5.19)$$

e) Ein spezieller Fall: der gewöhnliche geschaltete Poisson-Prozeß

Der in Gl. (5.18) angegebene Ausdruck gilt für beliebige Verteilungen von T_{A1}, T_{A2}, T_1 und T_2. Gemäß der Definition des geschalteten Poisson-Prozesses sind T_{A1} und T_{A2} negativ exponentiell verteilt. Weisen die Phasen T_1 und T_2 ebenfalls eine negativ exponentielle Verteilung auf, so erhält man den gewöhnlichen geschalteten Poisson-Prozeß:

$$F_1(t) = P\{T_1\le t\} = 1 - e^{-\omega_1 t}, \qquad (5.20a)$$

$$F_2(t) = P\{T_2\le t\} = 1 - e^{-\omega_2 t}. \qquad (5.20b)$$

Aufgrund der gedächtnislosen Eigenschaft der negativ exponentiellen Verteilung gemäß Gl. (4.2) bzw. Gl. (5.5), nach welcher die Vorwärts-Rekurrenzzeit durch die Zwischenankunftszeit ersetzt werden kann, erhält man

$$T_{mi}^v = \min\{T_{Ai}^v, T_i^v\} = \min\{T_{Ai}^v, T_i\} = \min\{T_{Ai}, T_i\} = T_{mi}, \qquad i = 1,2. \qquad (5.21a)$$

Hieraus folgt

$$\Phi_{mi}^v(s) = \Phi_{mi}(s) = LST\{F_{mi}(t) = P\{T_{mi}\le t\}\}, \qquad i = 1,2 \qquad (5.21b)$$

und

$$p_{Oi}^v = p_{Oi} \quad ; \quad p_{1i}^v = p_{1i} = 1 - p_{Oi}, \qquad i = 1,2. \qquad (5.21c)$$

Damit ergibt sich die Laplace-Stieltjes-Transformierte der Zwischenankunftszeit-Verteilung des gewöhnlichen geschalteten Poisson-Prozesses unter der Erneuerungsannahme gemäß Gl. (5.18) und Gl. (5.19):

$$\Phi(s) = 1 - \frac{s}{\lambda}\Phi^v(s),$$

mit

$$\Phi^v(s) = \frac{p_1\Phi_{m1}(s)(p_{11}+p_{O1}p_{12}\Phi_{m2}(s))+p_2\Phi_{m2}(s)(p_{12}+p_{O2}p_{11}\Phi_{m1}(s))}{1-p_{O1}p_{O2}\Phi_{m1}(s)\Phi_{m2}(s)}. \qquad (5.22)$$

Die Verteilungen für $T_{m1} = \min\{T_{A1}, T_1\}$ und $T_{m2} = \min\{T_{A2}, T_2\}$ sowie die Wahrscheinlichkeiten p_{11}, p_{12}, p_{O1}, p_{O2} werden im folgenden nacheinander hergeleitet.

Nach Gl. (5.9a), (5.20a) und der Notation gemäß Gl. (5.8) lautet die Verteilungsfunktion für die ZV $T_{m1} = \min\{T_{A1}, T_1\}$:

$$F_{m1}(t) = P\{T_{m1} \le t\} = 1 - P\{T_{m1} > t\}$$

$$= 1 - P\{T_{A1} > t\} . P\{T_1 > t\}$$

$$= 1 - (1 - F_{A1}(t)) . (1 - F_1(t))$$

$$= 1 - e^{-\lambda_1 t} e^{-\omega_1 t} = 1 - e^{-(\lambda_1+\omega_1)t}$$

oder nach der Laplace-Stieltjes-Transformation:

$$\Phi_{m1}(s) = LST\{F_{m1}(t)\} = \frac{\lambda_1 + \omega_1}{s + \lambda_1 + \omega_1} . \qquad (5.23a)$$

Analog erhält man

$$\Phi_{m2}(s) = LST\{F_{m2}(t)\} = \frac{\lambda_2 + \omega_2}{s + \lambda_2 + \omega_2} . \qquad (5.23b)$$

Mit der Betrachtung der bedingten Wahrscheinlichkeit

$$p_O(t) = P\{\text{keine Anforderungen in Phase i} \mid \text{Dauer der Phase i beträgt t}\}$$

$$= e^{-\lambda_i t} , \quad i = 1,2 ,$$

läßt sich mit der Verteilung der Phase 1 in Gl. (5.20a) die Wahrscheinlichkeit p_{O1}, daß in Phase 1 keine Anforderungen eintreffen, wie folgt berechnen:

$$p_{Oi} = \int_0^\infty p_O(t) f_i(t) dt = \int_0^\infty e^{-\lambda_i t} \omega_i e^{-\omega_i t} dt = \frac{\omega_i}{\omega_i+\lambda_i} , \; i = 1,2 \qquad (5.24a)$$

und

$$p_{1i} = 1 - p_{Oi} = \frac{\lambda_i}{\omega_i + \lambda_i} , \quad i = 1,2. \qquad (5.24b)$$

Werden die gewonnenen Größen aus Gl. (5.23ab) und Gl. (5.24ab) in Gl. (5.22) eingesetzt, so erhält man

$$\Phi^V(s) = \frac{1}{\omega_1+\omega_2} \cdot \frac{\lambda_1\omega_2(s+\omega_1+\omega_2+\lambda_2)+\lambda_2\omega_1(s+\omega_1+\omega_2+\lambda_1)}{(s+\lambda_1)(s+\lambda_2)+\omega_1(s+\lambda_2)+\omega_2(s+\lambda_1)} . \qquad (5.25)$$

Aus Gl. (5.19) und Gl. (5.25) ergibt sich die Laplace-Stieltjes-Transformation der Zwischenankunftszeiten:

$$\Phi(s) = \frac{1}{\lambda_1\omega_2+\lambda_2\omega_1} \cdot \frac{s(\lambda_1^2\omega_2+\lambda_2^2\omega_1)+(\lambda_1\lambda_2+\lambda_1\omega_2+\lambda_2\omega_1)(\lambda_1\omega_2+\lambda_2\omega_1)}{s^2+s(\lambda_1+\lambda_2+\omega_1+\omega_2)+\lambda_1\lambda_2+\lambda_1\omega_2+\lambda_2\omega_1} . \qquad (5.26)$$

Der Ausdruck für $\Phi(s)$ in Gl. (5.26) hat die Form

$$\Phi(s) = K \frac{s + a}{(s + s_1)(s + s_2)} ,$$

wobei

$$K = \frac{\lambda_1^2\omega_2 + \lambda_2^2\omega_1}{\lambda_1\omega_2 + \lambda_2\omega_1} \quad ; \quad a = \frac{\lambda_1\lambda_2 + \lambda_1\omega_2 + \lambda_2\omega_1}{K} \quad ;$$

$$s_{1/2} = \frac{1}{2}b \pm \frac{1}{2}\sqrt{b^2 - 4aK} \quad \text{mit } b = \lambda_1+\lambda_2+\omega_1+\omega_2.$$

Nach der Laplace-Rücktransformation erhält man die Verteilungsdichtefunktion der Zwischenankunftszeiten

$$f(t) = LT^{-1}\{\Phi(s)\} = \frac{K}{s_2-s_1}\left[(a-s_1)e^{-s_1 t} + (s_2-a)e^{-s_2 t}\right] . \qquad (5.27)$$

Der Mittelwert E[T] und der Variationskoeffizient c lassen sich aus Gl. (5.27) berechnen [57]:

$$E[T] = \frac{1}{\lambda} , \qquad (5.28a)$$

$$c^2 = \frac{E[T^2]}{E[T]^2} - 1 = 2\frac{\lambda(\lambda_1+\lambda_2+\omega_1+\omega_2-\lambda)}{\lambda_1\lambda_2+\lambda_1\omega_2+\lambda_2\omega_1} - 1 . \qquad (5.28b)$$

Wie zu erwarten ist, enthält der gewonnene Ausdruck in Gl. (5.25) die bekannten Ergebnisse der zwei Grenzprozesse, den Poisson-Prozeß und den unterbrochenen Poisson-Prozeß. Man erhält aus Gl. (5.25):

- Poisson-Prozeß: $\lambda_1 = \lambda_2 = \lambda$

$$\Phi(s) = \frac{\lambda}{s + \lambda} \qquad (5.29a)$$

- Unterbrochener Poisson-Prozeß: $\lambda_1 = 0$

$$\Phi(s) = \frac{\lambda_2(s + \omega_1)}{s^2 + s(\omega_1 + \omega_2 + \lambda_2) + \lambda_2\omega_1} \quad \text{(vgl. Kuczura [50]).} \qquad (5.29b)$$

Da diese Grenzprozesse die Erneuerungseigenschaft aufweisen, ist die Erneuerungsannahme für diese zwei Fälle exakt, wie aus den nachfolgend vorgestellten Ergebnissen zu ersehen ist.

f) Genauigkeit der Erneuerungsapproximation am Beispiel des Systems SPP/M/1-S

Die Genauigkeitsuntersuchung erfolgt hier anhand des Warteverlustsystems SPP/M/1-S, das aus einer Bedienungseinheit mit negativ exponentiell verteilter Bedienungsdauer und aus einer auf S begrenzten Wartekapazität besteht. Der geschaltete Poisson-Prozeß bildet dabei den Eingangsverkehrsstrom. Zur Illustration der Gültigkeitsbereiche der Erneuerungsannahme werden einige numerische Ergebnisse diskutiert. Die Systemanalyse wird ausführlich in [57] behandelt. Während die exakte Berechnung mit einem zweidimensionalen Markoff-Prozeß durchgeführt wird, lassen sich die

Ergebnisse bei der Erneuerungsapproximation durch die Analyse des Modells vom Typ GI/M/1-S mittels einer eingebetteten Markoff-Kette (s. Kap. 4.2.2) bestimmen.

Bild 5.9 zeigt die Genauigkeit der Erneuerungsapproximation am Beispiel der Blockierungswahrscheinlichkeit B. Dabei wird B als Funktion des Überlastfaktors $\gamma = \frac{\lambda_2}{\lambda}$ gezeigt. Wie erwartet, stimmen die exakten und die unter der Erneuerungsannahme gewonnenen Ergebnisse im Falle des Poisson-Prozesses ($\gamma=\gamma_{min}=1$) und des unterbrochenen Poisson-Prozesses ($\gamma=\gamma_{max}=2$) überein. Für die gewählten Parameter ($\Theta=1, n_0=2$) stellt die Erneuerungsannahme eine gute Näherung dar. Dies gilt nur für kleinere Werte von n_0, bei denen aufgrund der großen Schalthäufigkeit (vgl. Gl. 5.12) die Abhängigkeit des Prozesses von den Phasen gering ist.

Die Approximationsgenauigkeit ist auch in Bild 5.10 ersichtlich, in dem die mittlere Anzahl von Anforderungen im System als Funktion des Verkehrsangebotes ρ dargestellt wird ($\rho=\lambda h$; h: mittlere Bedienungsdauer der Anforderungen).

Die Vorteile der Prozeßbeschreibung mit der Erneuerungsannahme liegen darin, daß

- im Falle des gewöhnlichen geschalteten Poisson-Prozesses die Systemanalyse aufgrund der geschlossenen Form der Verteilungsfunktion entscheidend vereinfacht wird [57].
- im Falle der geschalteten Poisson-Prozesse mit allgemein verteilten Phasen (T_1 und T_2) eine Modellanalyse ermöglicht wird. Da der Eingangsprozeß mit einer Verteilungsfunktion approximativ beschreibbar ist (GI-Ankunftsprozeß nach Gl. 5.18), können abhängig vom Modell die aus der Warteschlangentheorie bekannten Analysemethoden angewendet werden.

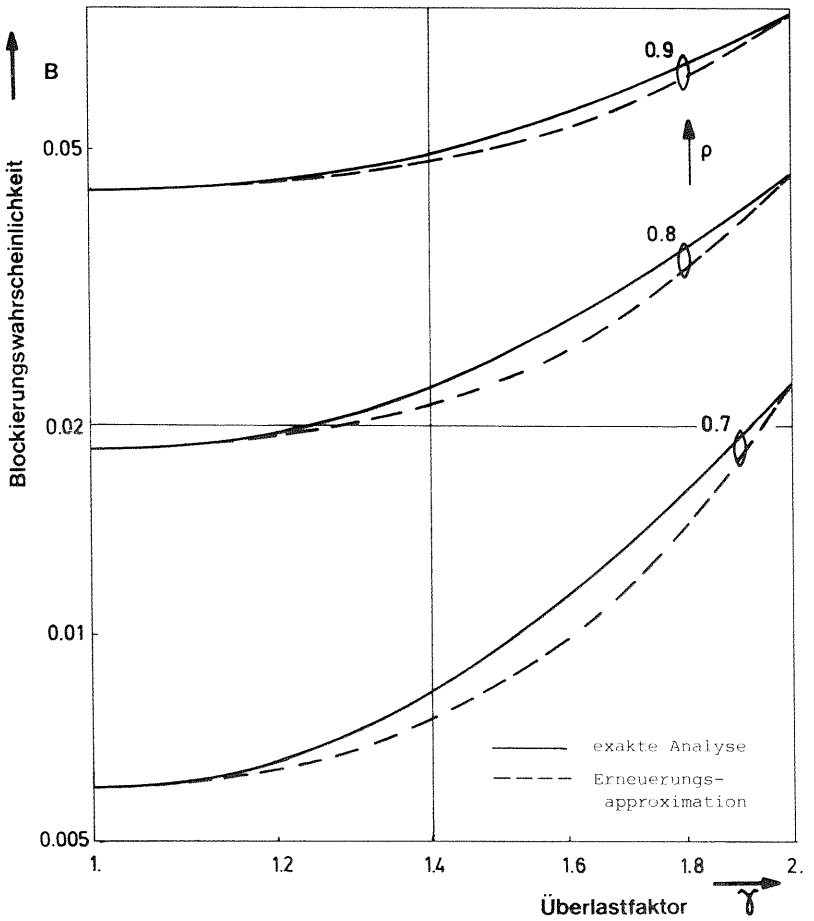


Bild 5.9 *Blockierungswahrscheinlichkeit des Warteverlustsystems SPP/M/1-S.*

Parameter : $S = 10$, $\Theta = 1$, $n_0 = 2$

$\gamma_{min} = 1 \rightarrow \lambda_1 = \lambda_2 = \lambda$ *Poisson-Prozeß*

$\gamma_{max} = 2 \rightarrow \lambda_1 = 0$ *unterbrochener Poisson-Prozeß*

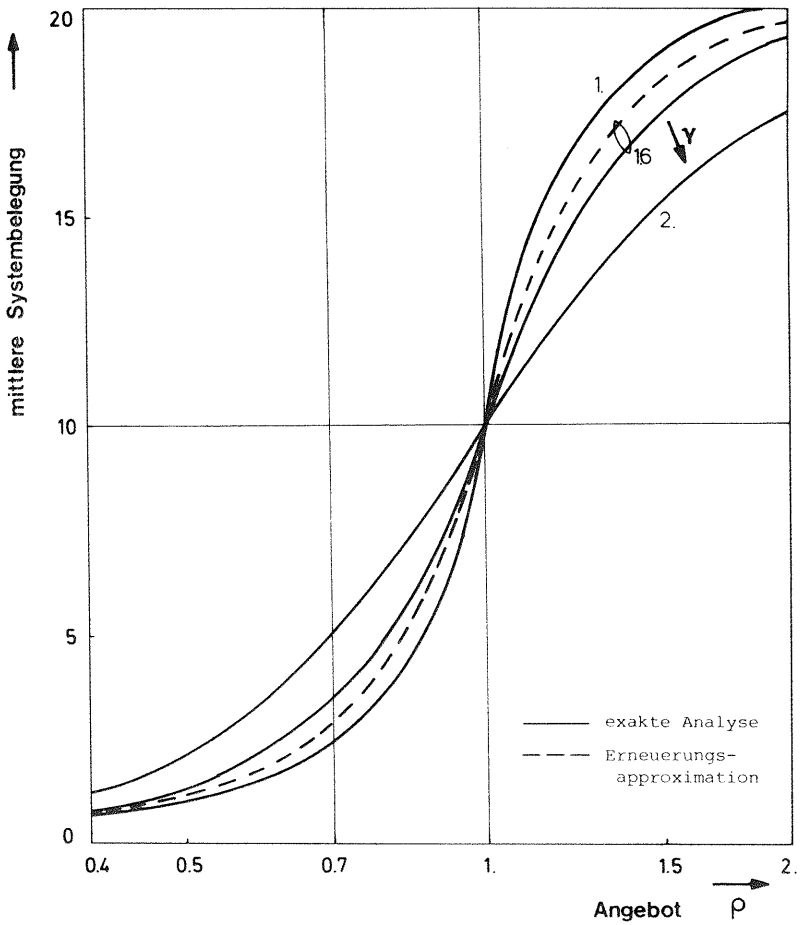


Bild 5.10 *Mittlere Anzahl von Anforderungen im Warteverlustsystem SPP/M/1-S.*

Parameter : $S = 20$, $\Theta = 1$, $n_0 = 10$

5.2 Rückwirkungsbehaftete Modelle zur Beschreibung von Überlastsituationen

Die Einflüsse des Teilnehmerverhaltens auf die Leistung eines Vermittlungssystems können nach zwei Hauptursachen unterteilt werden:

- Der Rufwiederholungseffekt, der zu ineffektiver Belegung der Vermittlungseinrichtung und daher zur Leistungssenkung führt
- Die gegenseitige Beeinflussung zwischen der Wartezeit der Teilnehmer, der Rufverarbeitungsdauer und der Rufkomplettierungsrate in einem Vermittlungssystem.

Im folgenden werden zwei Modellansätze vorgestellt, mit denen sich diese Zusammenhänge qualitativ und quantitativ untersuchen lassen. Mit Hilfe dieser Modelle kann die Entwicklung sowie das Zustandekommen von Überlastsituationen erklärt werden.

5.2.1 Modell für den Rufwiederholungseffekt

a) Allgemeines über Rufwiederholungsmodelle

Es findet sich in der Literatur eine Reihe von Arbeiten, die sich mit der Problematik der Rufwiederholung befassen [58-66]. In [61] wird unter Berücksichtigung der Geduld von Teilnehmern ein Verkehrsmodell für Rufwiederholung untersucht. Eine approximative Lösungsmethode für ein Rufwiederholungsmodell mit unendlicher Quellenzahl wird in [58] diskutiert. Während [64] und [66] Messungen bezüglich der Rufwiederholungseffekte vorstellen, werden in [62] und [63] detaillierte Modellbildungen sowie Simulationsstudien dargestellt. [59], [60] und [65] beschreiben Rufwiederholungsmodelle und diskutieren analytische Lösungsansätze.
Zur Ermittlung des Einflusses der Rufwiederholung wird hier ein verkehrstheoretisches Modell mit einer endlichen Anzahl von Verkehrsquellen vorgestellt. Mit der Annahme, daß die Wiederholwahrscheinlichkeit konstant bleibt, werden ein Algorithmus zur Modellanalyse entwickelt und numerische Ergebnisse vorgestellt.

b) Rufwiederholungsmodell mit endlicher Quellenzahl

Das in diesem Modellansatz benutzte Modell eines Teilnehmers geht aus Bild 5.11 hervor. Ein Teilnehmer bzw. eine Quelle wird durch folgende Zufallsvariablen charakterisiert:

T_{ID} Frei-Zeitdauer (idle time); Zeitspanne, in welcher sich die Quelle im Ruhezustand befindet. Die Quelle wird hier als Poisson-Quelle mit der Rate β betrachtet, d.h.

$$F_{ID}(t) = P\{T_{ID} \leq t\} = 1 - e^{-\beta t},$$
$$E[T_{ID}] = 1/\beta. \qquad \ldots(5.30)$$

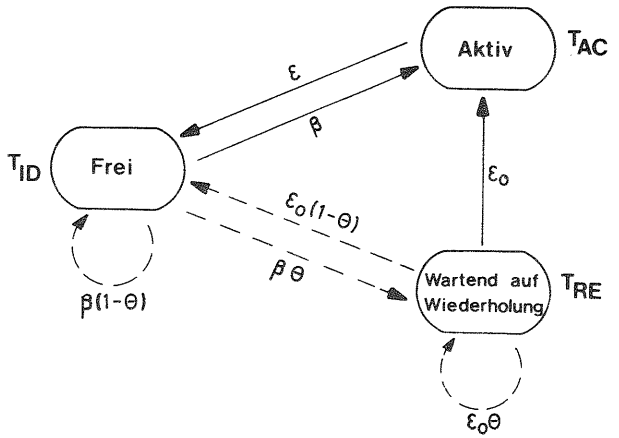


Bild 5.11 *Modell des Rufwiederholungsverhaltens eines Teilnehmers.*

T_{AC} Dauer der aktiven Zeit (active time). T_{AC} modelliert hier die Rufdauer, die hinsichtlich verschiedener Ruftypen bzw. Belegungsspektren mit einer negativ exponentiellen Verteilung beschrieben wird

$$F_{AC}(t) = P\{T_{AC} \leq t\} = 1 - e^{-\varepsilon t},$$
$$E[T_{AC}] = 1/\varepsilon. \qquad \ldots(5.31)$$

T_{RE} Rufwiederholabstand (inter-reattempt time); Zeitspanne zwischen Erstruf und erstem Folgeruf bzw. zwischen aufeinanderfolgenden Folgerufen. T_{RE} wird hier als unabhängige Zufallsvariable und mit einer negativ exponentiellen Verteilung modelliert

$$F_{RE}(t) = P\{T_{RE} \leq t\} = 1 - e^{-\varepsilon_0 t},$$
$$E[T_{RE}] = 1/\varepsilon_0. \qquad \ldots(5.32)$$

Die Übergänge zwischen den drei Grundzuständen einer Quelle sind in Bild 5.11 dargestellt. Sie sind davon abhängig, ob der Rufversuch (Erstruf oder Wiederholung) angenommen wird oder nicht. Der in den Übergangswahrscheinlichkeiten enthaltene Faktor Θ ist die Rufwiederholwahrscheinlichkeit. Diese Wahrscheinlichkeit hängt von der Geduld der Teilnehmer ab und ist i.a. abnehmend mit der Anzahl der Wiederholungen. Für den hier behandelten Modellansatz wird Θ als konstant angenommen, wobei in erster Linie der Effekt der Rufwiederholung bei endlicher Quellenzahl Gegenstand der Untersuchung ist. Das Gesamtmodell ist in Bild 5.12 dargestellt, dessen Komponenten im folgenden erläutert werden:

- Die endliche Anzahl q von Verkehrsquellen, mit denen Teilnehmer mit dem in Bild 5.11 beschriebenen Verhalten modelliert werden, erzeugen den Ankunftsverkehr. Die Verkehrsintensität ist von der Anzahl freier Teilnehmer bzw. Quellen abhängig.
- Die Anzahl n von Bedienungseinheiten, welche die Rufverarbeitung repräsentieren. Damit können für verschiedene Anwendungen des Modells Wahlaufnahmesätze, rufbezogene Datenblöcke oder Verbindungsleitungen modelliert werden. Die Bedienungsdauer wird hier als negativ exponentiell verteilte Zufallsvariable angenommen. Die dazugehörige Enderate einer Belegung ist ε.

Das Bedienungssystem wird als Verlustsystem betrieben, d.h. falls alle Bedienungseinheiten belegt sind, werden ankommende Rufversuche abgewiesen. Ein abgewiesener Ruf - Erstruf oder Rufwiederholung - wird mit der Wahrscheinlichkeit Θ wiederholt oder mit der komplementären Wahrscheinlichkeit $(1-\Theta)$ aufgegeben.

- Der Wiederholungsraum, in dem sich die auf eine Wiederholung wartenden Rufe aufhalten. Da die Anzahl der Quellen endlich ist, können maximal (q-n) Rufe gleichzeitig auf eine Wiederholung warten. Gemäß Gl. (5.32) besitzen die Warteplätze des Wiederholungsraumes die Enderate ε_0.

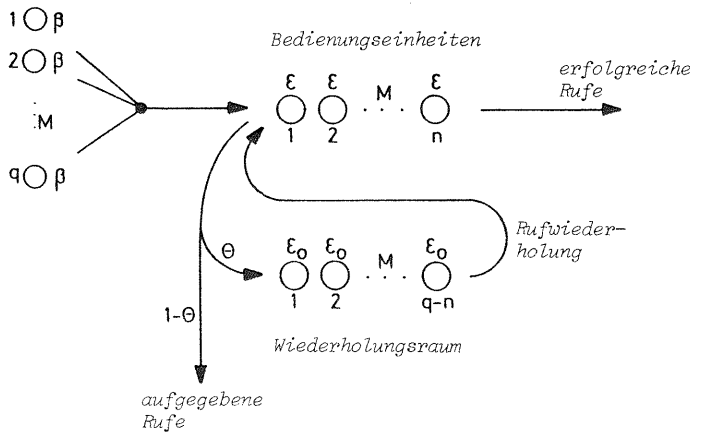


Bild 5.12 *Verkehrsmodell für Rufwiederholung mit endlicher Quellenzahl.*

c) Zustandsraum und Zustandsübergänge

Folgende Symbole werden für die Modellanalyse angewendet:

β	Anrufrate einer Verkehrsquelle im Freizustand
q	Anzahl der Verkehrsquellen
ε	Enderate einer Bedienungseinheit
n	Anzahl der Bedienungseinheiten
ε_o	Enderate eines Rufes im Wiederholungsraum
Θ	Rufwiederholwahrscheinlichkeit
$\rho_o = \frac{\beta q}{n}$	normiertes Verkehrsangebot
X	Zufallsvariable für die aktuelle Anzahl belegter Bedienungseinheiten
Z	Zufallsvariable für die aktuelle Anzahl der Rufe (Teilnehmer), die auf eine Wiederholung warten
$P(i,j) = P\{X=i, Z=j\}$	Zustandswahrscheinlichkeit dafür, daß i Rufe in Bedienung sind und j Rufe auf eine Wiederholung warten.

Mit den Zufallsvariablen X und Z läßt sich der Zustandsprozeß des Systems vollständig beschreiben. Da alle Modellkomponenten die Markoff-Eigenschaft aufweisen, kann im stationären Falle der Zustandsprozeß mit einem zweidimensionalen Markoff-Prozeß, dessen Zustandsübergangsdiagramm in Bild 5.13 dargestellt wird, beschrieben werden.

Betrachtet wird der Zustand $\{X=i, Z=j\}$ $(i < n)$, bei dem folgende Ereignisse zu einer Zustandsänderung führen können

- Ankunft eines Erstrufs: da sich in dem betrachteten Zustand (q-i-j) Quellen im Freizustand befinden, geht das System mit der Übergangswahrscheinlichkeitsdichte (ÜWD) $(q-i-j)\beta$ in den Zielzustand (i+1,j) über.

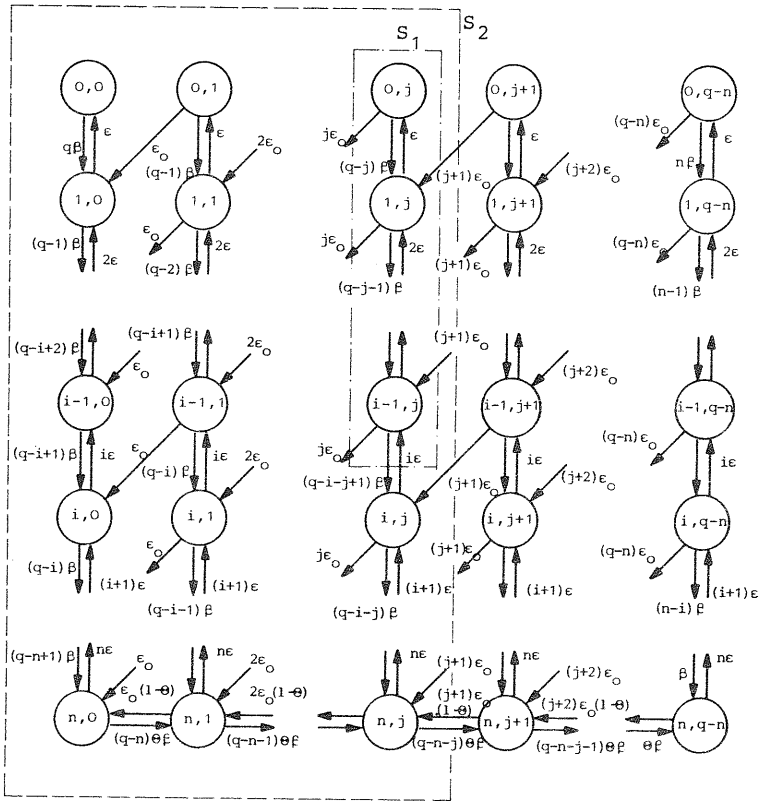


Bild 5.13 *Zustandsübergangsdiagramm des Modells für Rufwiederholung mit endlicher Quellenzahl.*

- Rufwiederholung: mit der ÜWD $j\varepsilon_o$ wird der Zielzustand (i+1,j-1) erreicht.
- Ruf-Bedienungsende: mit der ÜWD $i\varepsilon$ wird der Zielzustand (i-1,j) erreicht.

Eine Ausnahme bilden die Zustände (n,j), j=0,1,..,q-n, bei denen ein Erstruf bzw. eine Wiederholung zu folgenden Zustandsänderungen führt:

- Erstruf: mit der Wiederholwahrscheinlichkeit Θ wird der abgewiesene Ruf in den Wiederholungsraum transferiert; die ÜWD dafür ist $(q-n-j)\Theta\beta$.
- Rufwiederholung: mit der Wahrscheinlichkeit $(1-\Theta)$ wird eine abgewiesene Wiederholung aufgegeben; die ÜWD dafür ist $j\varepsilon_0(1-\Theta)$.

d) Analyse mittels numerischer Rekursion

Die Zustandswahrscheinlichkeiten werden mit einem rekursiven Algorithmus berechnet. Dafür werden zwei Grundbeziehungen benötigt, die aus dem Zustandsübergangsdiagramm gewonnen werden. Betrachtet werden zwei Makrozustände S_1 und S_2, die in Bild 5.13 gekennzeichnet werden. Die Übergangswahrscheinlichkeit zum Erreichen des Makrozustands S_1 lautet

$$(j+1)\varepsilon_0 \sum_{k=0}^{i-2} P(k,j+1) + i\varepsilon P(i,j), \tag{5.33a}$$

und die Übergangswahrscheinlichkeit zum Verlassen von S_1 errechnet sich zu

$$j\varepsilon_0 \sum_{k=0}^{i-1} P(k,j) + (q-i-j+1)\,\beta\, P(i-1,j). \tag{5.33b}$$

Befindet sich der Makrozustand S_1 im statistischen Gleichgewicht, so erhält man aus Gl. (5.33a) gemäß Gl. (5.33b) mit einfacher Umformung

$$i\varepsilon P(i,j) = j\varepsilon_0 \sum_{k=0}^{i-1} P(k,j)+(q-i-j+1)\beta P(i-1,j) - (j+1)\varepsilon_0 \sum_{k=0}^{i-2} P(k,j+1),$$

$$j = 0,1,\ldots,q-n,\ i = 0,1,\ldots,n, \tag{5.34}$$

wobei $P(i,j) = 0$ für $j > q-n$.

Analog erhält man für den Makrozustand S_2

$$(q-n-j)\Theta\beta P(n,j) = (j+1)\varepsilon_0 \sum_{k=0}^{n-1} P(k,j+1)+(j+1)(1-\Theta)\varepsilon_0 P(n,j+1),$$

$$j = 0,1,\ldots,q-n-1. \tag{5.35}$$

Die Gleichungen (5.34) und (5.35) bilden zusammen mit der Normierungsgleichung

$$\sum_{i=0}^{n} \sum_{j=0}^{q-n} P(i,j) = 1 \tag{5.36}$$

ein lineares Gleichungssystem zur Bestimmung der Zustandswahrscheinlichkeiten. Zur numerischen Berechnung wurde ein Algorithmus entwickelt, der aus folgenden Schritten besteht:

1) $P(0,q-n) = K_0$ setzen.
2) Für $P(i,q-n) = c_{i,q-n}K_0$ die Koeffizienten $c_{i,q-n}$, $i = 1,2,..,n$, mit Gl. (5.34) rekursiv berechnen. Die Wahrscheinlichkeiten $P(i,q-n)$, $i = 0,1,\ldots,n$, sind jetzt nur von $P(0,q-n) = K_0$ abhängig.
3) Spaltenindex $j = q-n-1$ setzen.
   $P(0,j) = K_1$ setzen.
4) Für $P(i,j) = u_{i,j}\,K_0 + v_{i,j}K_1$ die Koeffizienten $u_{i,j}$ und $v_{i,j}$ für $i = 1,2,\ldots,n$ mit Gl. (5.34) rekursiv berechnen. Man erhält schließlich die Zustandswahrscheinlichkeit

   $$P(n,j) = u_{n,j}\,K_0 + v_{n,j}\,K_1. \tag{5.37a}$$

   Andererseits läßt sich $P(n,j)$ mit Gl. (5.35) bestimmen zu

   $$P(n,j) = w_j\,K_0. \tag{5.37b}$$

   Aus Gl. (5.37ab) kann eine Beziehung zwischen K_1 und K_0 hergestellt werden

$$K_1 = \frac{w_j - u_{n,j}}{v_{n,j}} \quad K_0 = \alpha_j K_0. \tag{5.38}$$

5) Gemäß Gl. (5.38) die von K_0 und K_1 abhängigen Wahrscheinlichkeiten P(i,j) (j = 0,1,...,n) zwischennormieren, d. h. sie werden in K_0 wie folgt ausgedrückt:

$$P(i,j) = (u_{i,j} + \alpha_j v_{n,j})\, K_0 = c_{i,j}\, K_0,$$

$$i = 0,1,\ldots,n. \tag{5.39}$$

Nach diesem Schritt sind alle Wahrscheinlichkeiten P(i,k), i=0,1,...,n, k=j, j+1,..., q-n, nur noch von K_0 abhängig.

6) Die Schritte 3), 4) und 5) für j = q-n-2,...,1,0 wiederholen. Alle Zustandswahrscheinlichkeiten sind nun in K_0 ausgedrückt.

7) Zustandswahrscheinlichkeiten endnormieren. Dafür wird K_0 wie folgt berechnet:

$$\sum_{i=0}^{n} \sum_{j=0}^{q-n} P(i,j) = \sum_{i=0}^{n} \sum_{j=0}^{q-n} c_{i,j}\, K_0 = 1$$

$$\text{oder } K_0 = \left[\sum_{i=0}^{n} \sum_{j=0}^{q-n} c_{i,j} \right]^{-1}. \tag{5.40}$$

e) Systemcharakteristiken

Aus den Zustandswahrscheinlichkeiten, die mit Hilfe des beschriebenen numerischen Algorithmus gewonnen werden, lassen sich charakteristische Größen herleiten, mit denen der Einfluß des Rufwiederholungseffektes auf die Systemleistung untersucht wird.

Das Verkehrsgeschehen im Modell wird in Bild 5.14, in dem die Verkehrsströme im System illustriert werden, schematisch dargestellt. Folgende Indizierung wird für die Darstellung der mittleren Raten der Rufströme vorgenommen:

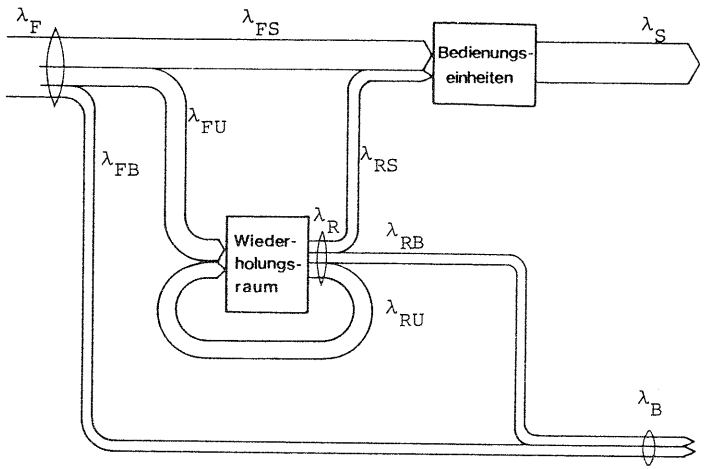


Bild 5.14 *Verkehrsflüsse im Rufwiederholungsmodell.*

F frischer Ruf bzw. Erstversuch (fresh call)

R Rufwiederholung (reattempt)

S erfolgreicher Ruf (successful call)

U abgewiesener Ruf, der wiederholt wird (unsuccessful call)

B abgewiesener Ruf, der aufgegeben wird (blocked call).

Nach dieser Indizierung ist λ_{FS} z.B. die mittlere Rate der Erstversuche, die sofort angenommen und bedient werden.

Die in Bild 5.14 dargestellten mittleren Verkehrsraten der Erstversuche lassen sich aus den Zustandswahrscheinlichkeiten folgendermaßen bestimmen:

$$\lambda_{FS} = \beta \sum_{i=0}^{n-1} \sum_{j=0}^{q-n} (q-i-j)\ P(i,j)\ ,$$

$$\lambda_{FU} = \Theta\beta \sum_{j=0}^{q-n} (q-n-j)\ P(n,j)\ ,$$

$$\lambda_{FB} = (1-\Theta)\beta \sum_{j=0}^{q-n} (q-n-j)\ P(n,j). \qquad \ldots(5.41)$$

Man erhält für die Wiederholungsverkehrsströme

$$\lambda_{RS} = \varepsilon_0 \sum_{i=0}^{n-1} \sum_{j=0}^{q-n} j\ P(i,j)\ ,$$

$$\lambda_{RU} = \Theta\varepsilon_0 \sum_{j=0}^{q-n} j\ P(n,j)\ ,$$

$$\lambda_{RB} = (1-\Theta)\varepsilon_0 \sum_{j=0}^{q-n} j\ P(n,j). \qquad \ldots(5.42)$$

Die gesamte Belastung des Systems setzt sich aus zwei Verkehrsströmen zusammen: die Erstversuche mit der mittleren Rate λ_F und die Rufwiederholungen mit der mittleren Rate λ_R:

$$\lambda_F = \lambda_{FS} + \lambda_{FU} + \lambda_{FB}\ ,$$

$$\lambda_R = \lambda_{RS} + \lambda_{RU} + \lambda_{RB}. \qquad \ldots(5.43)$$

Die Wahrscheinlichkeit dafür, daß ein Erstversuch abgewiesen wird, errechnet sich zu

$$B_F = \frac{\lambda_{FU} + \lambda_{FB}}{\lambda_F}\ . \qquad (5.44)$$

Charakteristisch für den Rufwiederholungseffekt ist die mittlere Anzahl der Versuche - einschließlich des Erstrufs -, die ein Ruf machen muß

$$\eta = \frac{\lambda_R + \lambda_F}{\lambda_F} = 1 + \frac{\lambda_R}{\lambda_F}\ . \qquad (5.45)$$

Betrachtet man nur die erfolgreichen Rufe, so muß ein erfolgreicher Ruf im Mittel η_F Versuche unternehmen :

$$\eta_F = \frac{\lambda_{FU} + \lambda_{FS} + \lambda_{RU} + \lambda_{RS}}{\lambda_{FS} + \lambda_{RS}} = \frac{\lambda_F + \lambda_R - \lambda_B}{\lambda_F - \lambda_B} = 1 + \frac{\lambda_R}{\lambda_F - \lambda_B}. \qquad (5.46)$$

Bild 5.15 zeigt die mittlere Anzahl der Rufversuche pro Ruf bei verschiedenen Systembelastungen $\rho_0 = \frac{\beta q}{n}$. ρ_0 stellt dabei lediglich das normierte Verkehrsangebot dar, da das tatsächliche Verkehrsangebot vom Systemzustand abhängig ist. Es ist ersichtlich, daß bei steigender Belastung ein Teilnehmer sehr oft Belegungsversuche unternehmen muß. Dieser Sachverhalt kann durch den "Schneeballeffekt" der Systemüberlastung begründet werden: steigende Belastung führt zu erhöhter Blockierungswahrscheinlichkeit; dies verursacht Rufwiederholungen, d. h. eine stärkere Systembelastung. Da die Rufe, die aufgegeben werden, im Mittel mit einer geringeren Anzahl von Versuchen verbunden sind, ist die Anzahl η_F bezüglich der erfolgreichen Rufe stets größer im Vergleich zu η. Falls alle abgewiesenen Rufe wiederholt werden (Θ=1), stimmen η und η_F wie erwartet überein.

Die Erhöhung der Rufblockierungswahrscheinlichkeit B_F infolge des Rufwiederholungseffektes zeigt Bild 5.16. Verglichen mit dem Fall, bei dem die Rufwiederholung nicht berücksichtigt wird (Θ=0), ist eine Erhöhung von B_F bei stärkerer Systembelastung für verschiedene Werte der Rufwiederholwahrscheinlichkeit Θ erkennbar.

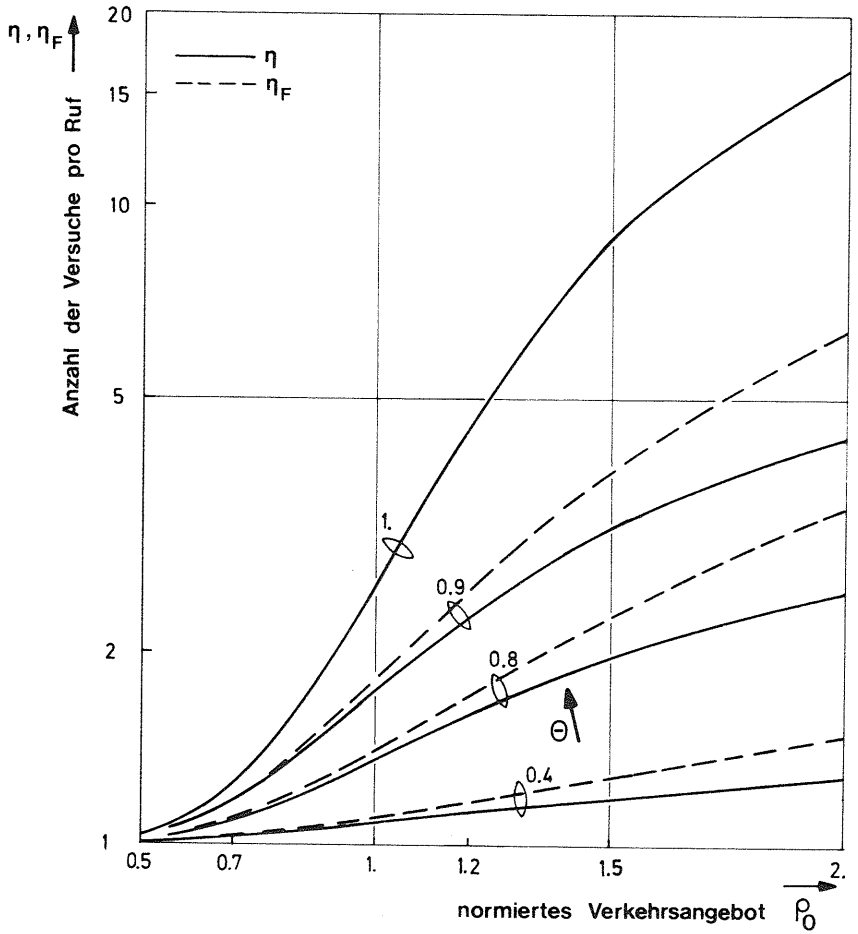


Bild 5.15 *Anzahl der Versuche pro Ruf für unterschiedliche Belastungen.*

Parameter : $\frac{\varepsilon_0}{\varepsilon} = 10$, $n = 10$, $q = 50$

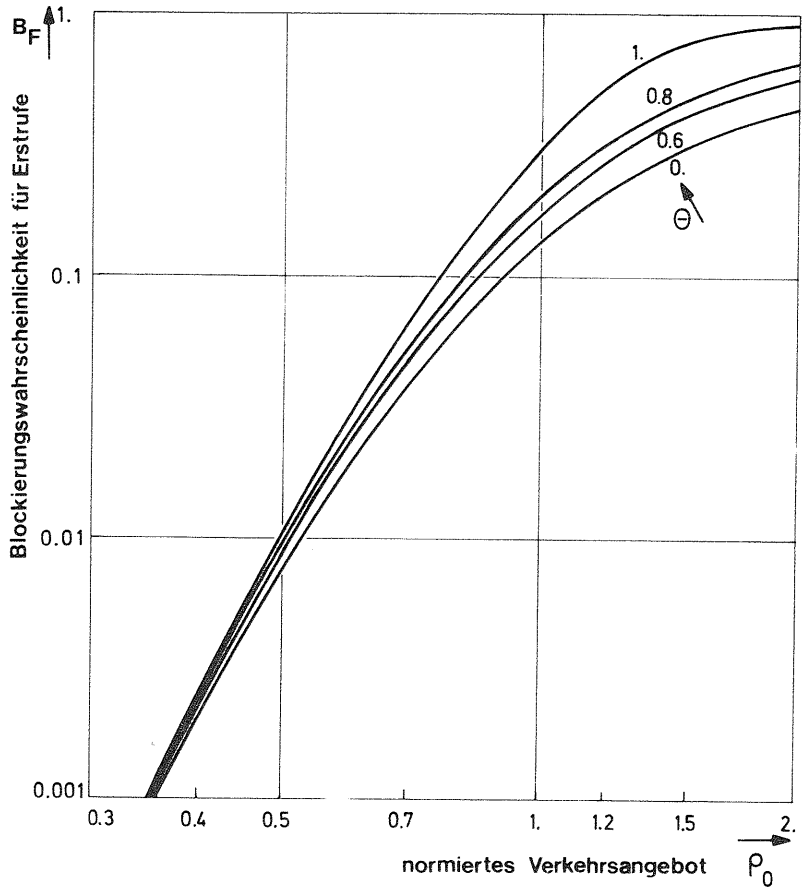


Bild 5.16 *Einfluß der Rufwiederholung auf die Rufblockierungswahrscheinlichkeit der Erstrufe.*

Parameter : $\frac{\varepsilon_0}{\varepsilon} = 10$, $n = 10$, $q = 50$

5.2.2 Überlastmodell mit wartezeitabhängiger Rufkomplettierung

a) Allgemeines

Das Ziel des hier beschriebenen Modellansatzes ist es, die Zusammenhänge zwischen der Wartezeit-Charakteristik von Teilnehmern, der Rufverarbeitungsdauer und der Rufkomplettierungsrate eines Vermittlungssystems zu erfassen. Die Wechselwirkungen zwischen diesen Komponenten werden vom Teilnehmerverhalten stark beeinflußt und führen häufig zu einer Verringerung der Leistungsfähigkeit eines Vermittlungssystems, besonders in Überlastsituationen.

Werden Teilnehmer, bedingt durch einen momentanen Systemengpaß, langen Wartezeiten ausgesetzt (z.B. bei langem Wähltonverzug), sind fehlerhafte Reaktionen auf der Teilnehmerseite wahrscheinlicher (z.B. Wahlbeginn vor dem Wählton), so daß Rufe häufiger aufgegeben bzw. unterbrochen werden. Die Folge ist, daß der Vermittlungsrechner einen Teil seiner Kapazität zur Verarbeitung von Steuerungsaufrufen aufwendet, die zu keiner Rufkomplettierung führen. Bei den nichtkomplettierten Rufen ist die Anzahl der Teilrufe bzw. Steuerungsaufrufe pro Ruf geringer als bei den erfolgreich verarbeiteten Rufen. Infolgedessen kann ein Zusammenhang zwischen der Wartezeit eines Rufes bis zur Rufaufnahme und seiner Verarbeitungsdauer hergestellt werden.

Die Abhängigkeit der Bedienungszeit von der Wartezeit wird in einigen Studien berücksichtigt [33,67,68,69], wobei in erster Linie stationäre Untersuchungen zu finden sind. In [33] wird anhand des M/M/1-Wartesystems die wartezeitabhängige Rufverarbeitungszeit diskutiert. Das Wartesystem M/G/1 mit wartezeitabhängiger Bedienungszeit wird in [67,68,69] untersucht. Während [67] den Sonderfall M/M/1 mit einer Wiener-Hopf-Integralgleichung behandelt, wird in [69] die Methode der eingebetteten Markoff-Kette angewendet. Eine spezielle Verteilungsfunktion für die Bedienungszeit wird in [68] untersucht.

Das in diesem Kapitel entwickelte Modell wird hinsichtlich des instationären Verhaltens untersucht, wobei die dynamische Entwicklung einer Überlastsituation ermittelt wird. Als Eingangsprozeß werden impulsförmige Überlastmuster nach dem in Gleichung (4.20b) angegebenen verallgemeinerten Poisson-Prozeß verwendet.

b) Modellbeschreibung

Der vorgestellte Modellansatz beruht auf einer Diskretisierung der Rufverarbeitungszeit (vgl. [70]). Die Rufverarbeitung setzt sich aus Verarbeitungszeiten einzelner Steuerungsaufrufe zusammen. Deshalb kann die kontinuierliche Zufallsvariable für die Rufverarbeitung durch eine diskrete Zufallsvariable für die Anzahl der Steuerungsaufrufe bzw. Teilrufe ersetzt werden. Die Verarbeitungszeit von Teilrufen, die im weiteren auch Phasen genannt werden, kann als unabhängig von der Wartezeit und vom Systemzustand angesehen werden.

Diese Überlegungen führen zu einem Verkehrsmodell vom Typ $M^{[x]}/G/1$, bei dem der Gruppenankunftsprozeß der Phasen zustandsabhängig ist. Wie später noch erläutert wird, beinhaltet die Zustandsabhängigkeit auch die zu modellierende Wartezeitabhängigkeit.

Werden alle Arten von Steuerungsaufrufen betrachtet, so kann für die Phasendauer T_{PH} eine negativ exponentielle Verteilung angenommen werden :

$$F_{PH}(t) = 1 - e^{-\mu t}, \tag{5.47}$$

mit $E[T_{PH}] = \frac{1}{\mu}$.

Diese Annahme wird durch eine Untersuchung in [70] unterstützt, in der gezeigt wird, daß durch eine ausreichend feine Diskretisierung der Rufverarbeitungszeit der Einfluß der Phasenverteilung vernachlässigbar klein bleibt.

Betrachtet wird nun ein Testruf, der zum Ankunftszeitpunkt eine Anzahl k der noch zu verarbeitenden Phasen im System antrifft. Die

Wartezeit W des Rufes entspricht einer Erlang-Verteilung k-ter Ordnung:

$$P\{W \leq t\} = E_k(t) = 1 - e^{-\mu t} \sum_{i=0}^{k-1} \frac{(\mu t)^i}{i!} \quad , \; t \geq 0. \tag{5.48}$$

Je länger die Wartezeit des Testrufs ist, desto geringer wird die Wahrscheinlichkeit zur Komplettierung und desto kürzer ist die Dauer der Rufverarbeitung. Sei j die Anzahl der Phasen, die ein Testruf für seine Verarbeitung benötigt, so ist j kleiner bei nichtkomplettierten Rufen im Vergleich zu erfolgreichen Rufen. Diese Feststellung erlaubt es, einen Zusammenhang zwischen k und j herzustellen, abhängig von der Wartezeitverteilung in Gl.(5.48). Dies führt zu einer zustandsabhängigen Zufallsvariable $G^{(k)}$. Ein Testruf, der zum Ankunftszeitpunkt k Phasen angetroffen hat, wird mit der Wahrscheinlichkeit $g_j^{(k)} = P\{G^{(k)}=j\}$ die Anzahl j der Verarbeitungsphasen generieren.

Ein Ruf mit j Phasen wird mit der Wahrscheinlichkeit c_j erfolgreich verarbeitet. Diese bedingte Rufkomplettierungswahrscheinlichkeit ist charakteristisch für das Teilnehmerverhalten. Da ein erfolgreicher Ruf normalerweise eine längere Rufaufbauphase benötigt, wird im Modell die bedingte Komplettierungswahrscheinlichkeit c_j größer bei höherer Anzahl der generierten Phasen.

Die Modellelemente des $M^{[X]}/M/1$ - Systems mit zustandsabhängiger Gruppenankunft werden in folgenden Punkten zusammengefaßt :

- Rufankunftsprozeß ist ein verallgemeinerter Poisson-Prozeß mit der zeitabhängigen Rate $\lambda(t)$ (vgl.Gl. 4.20a,b).

- Ein Ruf, der zum Ankunftszeitpunkt k Phasen im System antrifft, wird mit der Wahrscheinlichkeit $g_j^{(k)}$ j Phasen zur Verarbeitung erzeugen.

- Rufe, die j Phasen generieren, werden mit der Wahrscheinlichkeit c_j komplettiert.

Ferner werden folgende Aspekte des Teilnehmerverhaltens im Modell berücksichtigt :

- Analog der Teilrufbetrachtung hat die Anzahl der Phasen pro Ruf eine untere Grenze N_0 und eine obere Grenze N_1, die von der Teilruferzeugung und deshalb von der Systemstruktur abhängig sind :

$$g_j^{(k)} = 0 \quad \text{für } j > N_1 \text{ oder } j < N_0. \tag{5.49}$$

- Überschreitet die Anzahl k der Phasen bei der Rufankunft eine Grenze k_0 $(k \geq k_0)$, so bleibt die Verteilung der Phasenanzahl $G^{(k)}$ unverändert :

$$g_j^{(k)} = g_j^{(k_0)} \text{ für } k \geq k_0 \; , \; N_0 \leq j \leq N_1. \tag{5.50}$$

Durch diese Annahmen wird die Modellanalyse entscheidend vereinfacht.

c) Parameterfestlegung

Folgende Symbole werden in diesem Modell angewendet :

λ Rufankunftsrate.

$h = \frac{1}{\mu}$ mittlere Bedienungszeit einer Phase.

$G^{(k)}$ Zufallsvariable (ZV) für die Anzahl der Phasen eines Rufes, der zum Ankunftszeitpunkt k Phasen im System antrifft;

$G^{(k)} = G^{(k_0)}$ für $k \geq k_0$ nach Gl.(5.50).

$g_j^{(k)} = P\{G^{(k)}=j\}$

Zustandsabhängige Gruppenverteilung.

$\rho_0 = \lambda h \, E[G^{(k_0)}]$

normiertes Rufangebot.

$X(t)$ ZV für die aktuelle Anzahl der Phasen im System zum Zeitpunkt t.

$P_i(t) = P\{X(t)=i\}$

zeitabhängige (instationäre) Zustandswahrscheinlichkeit.

$P_i = \lim_{t\to\infty} P_i(t)$

stationäre Zustandswahrscheinlichkeit.

c_j bedingte Komplettierungswahrscheinlichkeit.

Die nachfolgend beschriebene Analyse ist unabhängig von der Wahl der Verteilung für die ZV $G^{(k)}$ und von den bedingten Komplettierungswahrscheinlichkeiten c_j. Zur Durchführung der Untersuchung hinsichtlich der Systemreaktion in Überlastfällen werden für $g_j^{(k)}$ und c_j Festlegungen gemacht, die das Teilnehmerverhalten approximativ darstellen.

Die zustandsabhängige Gruppenverteilung $g_j^{(k)}$ wird in Anlehnung an die Erlang-k Verteilung der Wartezeit der Rufe gemäß Gl. (5.48) wie folgt bestimmt:

$$k=0 \quad \begin{cases} g_{N_1}^{(0)} = 1 \\ g_j^{(0)} = 0 \quad \text{sonst} , \end{cases}$$

$$0<k<k_0 \quad \begin{cases} g_{N_0}^{(k)} = 1 - E_k[(N_1-N_0)h] \\ g_j^{(k)} = E_k[(N_1-j+1)h] - E_k[(N_1-j)h]; N_0<j\le N_1 , \end{cases}$$

$$k=k_0 \quad \begin{cases} g_{N_0}^{(k_0)} = 1 \\ g_j^{(k_0)} = 0 \quad \text{sonst.} \end{cases} \qquad \ldots(5.51)$$

Die Wahl von $g_j^{(k)}$ wird in Bild 5.17 näher betrachtet, in dem die Mittelwerte $E[G^{(k)}]$ als Funktion von k aufgetragen werden. Die Erwartungswerte $E[G^{(k)}]$ werden dabei folgendermaßen berechnet ($k<k_0$):

$$E[G^{(k)}] = \sum_{j=N_0}^{N_1} j g_j^{(k)} = N_0 + \sum_{j=N_0}^{N_1} (j-N_0) g_j^{(k)}$$

$$= N_0 + \sum_{j=N_0+1}^{N_1} g_j^{(k)} \sum_{i=1}^{j-N_0} 1 = N_0 + \sum_{i=1}^{N_1-N_0} \sum_{j=N_0+i}^{N_1} g_j^{(k)}$$

$$= N_0 + \sum_{i=1}^{N_1-N_0} \sum_{j=N_0+i}^{N_1} (E_k[(N_1-j+1)h] - E_k[(N_1-j)h])$$

$$= N_0 + \sum_{i=1}^{N_1-N_0} E_k[(N_1-N_0-i+1)h] = N_0 + \sum_{i=1}^{N_1-N_0} E_k[ih]. \qquad (5.52)$$

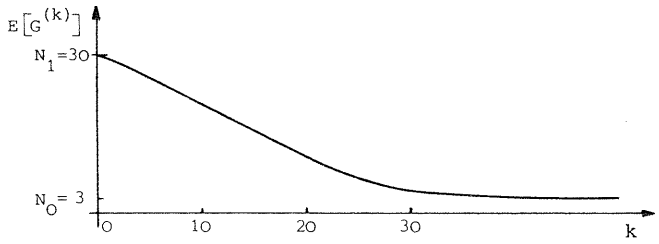


Bild 5.17 *Mittelwert der wartezeitabhängigen Rufverarbeitungsdauer.*

Die Koordinaten k und $E[G^{(k)}]$ in Bild 5.17 können als die mittlere Wartezeit und die mittlere Rufdauer interpretiert werden. Mit der Annahme von $G^{(k)}$ gemäß Gl. (5.51) wird die Wartezeitabhängigkeit approximativ so beschrieben, daß Rufe mit langer Wartezeit eine geringere Anzahl von Teilrufen zur Verarbeitung einbringen. Die bedingten Komplettierungswahrscheinlichkeiten c_j, $j = N_0,\ldots,N_1$, werden unter Berücksichtigung der Tatsache festgelegt, daß im Vergleich zu erfolgreichen Rufen nicht-komplettierte Rufe in der Regel kürzere Verarbeitungszeiten benötigen

$$c_j = \begin{cases} \gamma + (1-\gamma)\dfrac{j-N_0}{N_{LIM}-N_0} , & N_0\le j\le N_{LIM} , \\ 1 , & N_{LIM}\le j\le N_1 , \\ 0 , & \text{sonst.} \end{cases} \qquad (5.53)$$

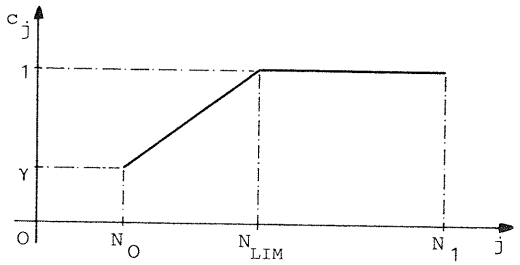


Bild 5.18 *Bedingte Rufkomplettierungswahrscheinlichkeiten.*

Die in Gl. (5.53) angegebene bedingte Komplettierungswahrscheinlichkeit c_j stellt zusammen mit der zustandsabhängigen Verteilung $g_j^{(k)}$ der Phasenanzahl den wartezeitbezogenen Teil des Teilnehmerverhaltens dar.

d) Stationäre und instationäre Analyse

Der Zustandsprozeß wird mit der zeitabhängigen Zufallsvariable X(t) gekennzeichnet, welche die aktuelle Anzahl der Phasen im System zum Zeitpunkt t beinhaltet. Betrachtet man das System zu einem infinitesimal kurzen Zeitintervall (t,t+dt), so kann für den Zustandsprozeß das in Bild 5.19 dargestellte Zustandsübergangsdiagramm entwickelt werden. Am Beispiel des Zustands i wird dieses Diagramm verdeutlicht:

- Ankunft eines Rufes mit j Phasen: Die Übergangswahrscheinlichkeitsdichte (ÜWD) dafür ist $\lambda(t)g_j^{(k)}$ $(N_0 \leq j \leq N_1)$, so daß nur die Zustände $i+N_0$ bis $i+N_1$ vom Zustand i aus erreicht werden können.
- Bearbeitungsende einer Phase: Mit der ÜWD μ geht das System in den Zustand i-1 über (i>0).

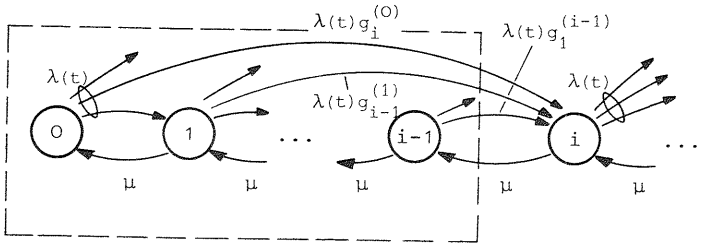


Bild 5.19 *Zustandsübergangsdiagramm.*

Für den Zustand i kann gemäß Gl. (4.8a) die Zustandsgleichung im instationären Falle angegeben werden

$$\frac{d}{dt}P_i(t) = -(\mu+\lambda(t))\,P_i(t)+\lambda(t)\sum_{n=0}^{i-1}P_n(t)g_{i-n}^{(n)}+\mu P_{i+1}(t),$$

$$i = 1,2,\ldots, \qquad (5.54a)$$

und für den Zustand i=0 gilt

$$\frac{d}{dt}P_0(t) = -\lambda(t)P_0(t) + \mu P_1(t). \qquad (5.54b)$$

Die Gleichungen (5.54a,b) bilden ein Differenzen-Differential-Gleichungssystem, dessen Auflösung die transienten Zustandswahrscheinlichkeiten liefert. Die Auflösung erfolgt hier numerisch, wobei das Verfahren nach Runge-Kutta [75] angewendet wird. Dazu

muß der unendliche Zustandsraum in Bild 5.19 in geeigneter Weise abgeschnitten werden, indem sehr kleine Wahrscheinlichkeiten zu Null gesetzt werden.

Aus den zeitabhängigen Zustandswahrscheinlichkeiten lassen sich die Systemcharakteristiken berechnen, z.B. die mittlere Anzahl der Phasen im System

$$E[X(t)] = \sum_{k=1}^{\infty} k P_k(t) \tag{5.55}$$

und die zeitabhängige Rufkomplettierungswahrscheinlichkeit

$$C(t) = \sum_{k=0}^{\infty} P_k(t) \sum_{j=N_0}^{N_1} c_j g_j^{(k)}. \tag{5.56}$$

Die transiente Rufkomplettierungsrate errechnet sich zu

$$Y(t) = \lambda(t) C(t) \ , \tag{5.57}$$

und die normierte Rufkomplettierungsrate lautet

$$Y_0(t) = Y(t) h E[G^{(k_0)}]. \tag{5.58}$$

Sind zu einem bestimmten Zeitpunkt, z.B. t=0, die sog. Start-Wahrscheinlichkeiten bekannt, so können durch die numerische Auflösung des Gleichungssystems nach Gl. (5.54a,b) die charakteristischen Größen berechnet werden. Falls der zu untersuchende instationäre Prozeßabschnitt einem stationären Prozeßabschnitt folgt, so wird die Start-Verteilung zweckmäßigerweise mit einer stationären Analyse ermittelt. Im folgenden werden die wesentlichen Schritte der stationären Modellanalyse [70] beschrieben.

Betrachtet man in Bild 5.19 den gekennzeichneten Makrozustand S, der sich im statistischen Gleichgewicht befindet, so lautet die stationäre Zustandsgleichung :

$$\mu P_i = \lambda \sum_{k=0}^{i-1} P_k \sum_{j=i-k}^{\infty} g_j^{(k)} \quad ; \ i>0. \tag{5.59}$$

Durch eine Summation über i erhält man mit Gl. (5.59):

$$\mu \sum_{i=1}^{\infty} P_i = \lambda \sum_{i=1}^{\infty} \sum_{k=0}^{i-1} P_k \sum_{j=i-k}^{\infty} g_j^{(k)} = \lambda \sum_{k=0}^{\infty} \sum_{i=k+1}^{\infty} \sum_{j=i-k}^{\infty} P_k g_j^{(k)}$$

$$= \lambda \sum_{k=0}^{\infty} P_k \sum_{j=1}^{\infty} g_j^{(k)} \sum_{i=k+1}^{k+j} 1 = \lambda \sum_{k=0}^{\infty} P_k E[G^{(k)}],$$

oder mit der Festlegung für $G^{(k)}$ in Gl. (5.50):

$$\mu(1-P_0) = \lambda \sum_{k=0}^{k_0-1} P_k E[G^{(k)}] + \lambda E[G^{(k_0)}] \sum_{k=k_0}^{\infty} P_k \ ,$$

$$1-P_0 = \rho_0 \sum_{k=0}^{k_0-1} P_k \frac{E[G^{(k)}]}{E[G^{(k_0)}]} - \rho_0 \sum_{k=0}^{k_0-1} P_k + \rho_0.$$

Mit der Normierung

$$P_k = \sigma_k P_0 \tag{5.60}$$

erhält man schließlich

$$P_0 = (1-\rho_0) \left[1 + \rho_0 \sum_{k=0}^{k_0-1} \sigma_k \left(\frac{E[G^{(k)}]}{E[G^{(k_0)}]} - 1 \right) \right]^{-1} . \tag{5.61}$$

Die Gleichungen (5.59) und (5.61) führen zu einem numerischen Algorithmus zur Bestimmung stationärer Zustandswahrscheinlichkeiten [70]:

1) $P_0^* = 1$ setzen.
2) mit Gl. (5.59) P_k^* bzw. σ_k, $1 \leq k \leq k_0 - 1$, rekursiv berechnen.
3) P_0 nach Gl. (5.61) ermitteln.
4) mit der gewonnenen Wahrscheinlichkeit P_0 die weiteren Wahrscheinlichkeiten P_k, $k > 0$, nach Gl. (5.59) rekursiv berechnen.

Die charakteristischen Größen werden für den stationären Fall aus den Zustandswahrscheinlichkeiten gemäß Gl. (5.55-5.58) berechnet. Sie bilden ebenfalls die Startwerte für die in diesem Kapitel durchgeführten instationären Untersuchungen.

e) Instationäre Systemantwort auf kurzzeitige Überlastimpulse

Liegt ein impulsförmiges Überlastmuster am Eingang des Systems, so kann die Systemantwort, dargestellt durch die zeitabhängige Rufkomplettierungsrate Y(t), registriert werden. In dieser Weise wird das Zustandekommen einer Überlastsituation untersucht. Diese Denkweise ist der Ermittlung von Einschwingvorgängen in der klassischen Systemtheorie ähnlich. Bild 5.20 zeigt die dynamische Systemantwort auf ein dreieckförmiges Überlastmuster, das einer stationären Grundlast ρ_{OS} überlagert ist. Die kurzzeitige Überlastung wird durch die Überlastdauer T_p und die Fläche F, die dem Überlast-Verkehrsvolumen entspricht, charakterisiert. Zu Beginn der Verkehrserhöhung steigt die Rufkomplettierungsrate, da das System noch nicht ausgelastet ist. Mit zunehmender Belastung wird das System voller, die Wartezeit länger, so daß die Rufkomplettierungsrate abnimmt. Normalisiert sich die Belastung, so braucht das System eine gewisse Zeit, um den noch vorhandenen Überlastverkehr abzubauen und den ursprünglichen stationären Zustand wieder herzustellen. Während dieses Intervalls bleibt die Komplettierungsrate zeitweise unter dem stationären Wert, was eine Leistungssenkung des Systems bedeutet.

Die Eigenschaften der Systemantwort können deutlicher erkannt werden, indem die Entwicklung des Überlastfalls durch Trajektorien dargestellt wird. Aus Bild 5.21 ist ersichtlich, daß der instationäre Verlauf der Systemreaktion sehr stark von der stationären Durchsatzkurve abweicht in Abhängigkeit von der Überlastdauer. Hält die Überlast sehr lange an ($T_p \to \infty$), so geht die Trajektorie in die stationäre Kurve über. Für die Darstellung wurden in Bild 5.21 unterschiedliche Werte für T_p ausgewählt, womit der Unterschied der Systemreaktion auf kurzzeitige und langzeitige Überlastung verdeutlicht wird. Diese Erkenntnis aus der instationären Betrachtung ist sehr aufschlußreich zur Erfassung der Überlastsituation. Dabei können systemspezifische Zeitkonstanten ermittelt werden, welche für die Festlegung der Aktivierungszeitpunkte geeigneter Überlastabwehrmaßnahmen herangezogen werden können.

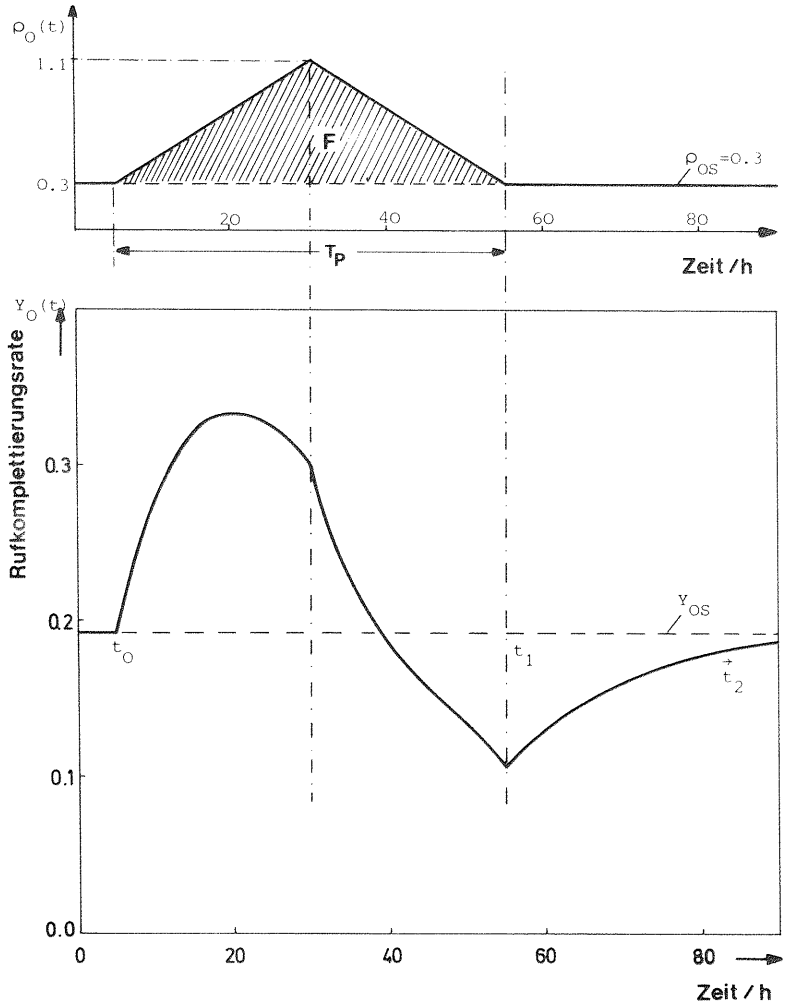


Bild 5.20 *Systemantwort auf einen dreieckförmigen Überlastimpuls.*

Parameter : $N_O = 3$ $N_{LIM} = 4$
$N_1 = 30$ $\gamma = 0.1$

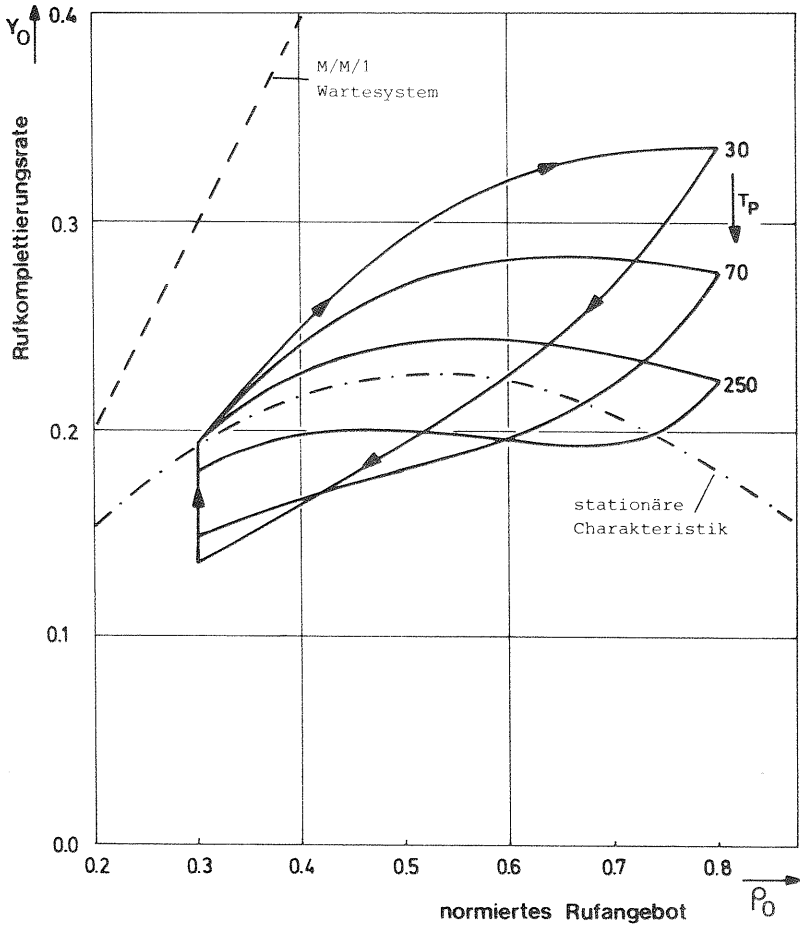


Bild 5.21 *Trajektorien der transienten Systemantwort auf dreieckförmige Überlastimpulse unterschiedlicher Dauer.*

Parameter : $N_O = 3$ $N_{LIM} = 4$
$N_1 = 30$ $\gamma = 0.1$

Um einen Vergleich der Systemantworten auf unterschiedliche Lastmuster zu ermöglichen, wird hier ein Gütefaktor für die Systemleistung in Überlastsituationen eingeführt, der die gesamte instationäre Reaktion erfaßt:

$$C_O = \frac{\text{komplettierter Anteil des Überlast-Verkehrs}}{\text{Überlast-Verkehrsvolumen}} = \frac{F_C}{F}. \quad (5.62)$$

Der Gütefaktor C_O kann hier als die Komplettierungswahrscheinlichkeit des zusätzlich eintreffenden Verkehrsvolumens interpretiert werden. Während F (vgl. Bild 5.20) das Überlast-Verkehrsvolumen darstellt

$$F = \int_{t_O}^{t_1} (\rho_O(t) - \rho_{OS})\, dt\ , \quad (5.63)$$

kann F_C wie folgt errechnet werden

$$F_C = \int_{t_O}^{t_2} (Y_O(t) - Y_{OS})dt = \int_{t_O}^{t_2} Y_O(t)dt - Y_{OS}(t_2 - t_O). \quad (5.64)$$

Aus Gl. (5.62), (5.63) und (5.64) ergibt sich

$$C_O = \frac{\int_{t_O}^{t_2} (Y_O(t) - Y_{OS})dt}{\int_{t_O}^{t_1} (\rho_O(t) - \rho_{OS})dt}\ . \quad (5.65)$$

In den Gleichungen (5.63-65) werden folgende Formelzeichen benutzt (Bild 5.20)

ρ_{OS}: stationäre Last (stationäres, normiertes Rufangebot)

Y_{OS}: Rufkomplettierungsrate bei der stationären Last ρ_{OS}

t_O : Beginn der Überlastung

t_1 : Ende der Überlastung

t_2 : Ende der Überlastsituation. Dies ist der Zeitpunkt, von dem an der stationäre Zustand wieder erreicht wird, d.h. die Abweichung vom stationären Verlauf vernachlässigbar klein wird.

In dem Ausdruck für F_C wird zunächst das Integral über alle komplettierten Rufe während der Überlastsituation gebildet. Davon wird der theoretische Anteil der komplettierten Rufe abgezogen, welcher der Grundlast (stationäre Last) entstammt.

Bild 5.22 zeigt die Abhängigkeit der Systemleistung von den Parametern der dreieckförmigen Überlast-Impulse, wobei die Komplettierungswahrscheinlichkeit C_O des Überlast-Verkehrs untersucht wird. Nimmt das Überlast-Verkehrsvolumen F zu, so kann ein Leistungsabfall des Systems festgestellt werden. Diese Leistungsminderung ist überdies sehr stark von der Überlastdauer T_p abhängig. Bei gleichem Überlast-Verkehrsvolumen, d. h. bei einem konstanten Wert für F, ist die Senkung der Systemleistung um so geringer, je länger die Überlastdauer T_p ist. Der Grund dafür liegt darin, daß bei langen, jedoch amplitudenmäßig nicht sehr starken Überlast-Impulsen das System größere Reserve-Kapazität zur Verarbeitung des zusätzlich angebotenen Verkehrs besitzt.

Die in der Gl. (5.65) angegebene Definition für den Gütefaktor C_O der Überlastverarbeitung kann allgemein auf alle impulsförmigen Überlastmuster angewendet werden [56]. Dies ermöglicht quantitative Untersuchungen hinsichtlich der Leistungsfähigkeit des Systems bei zeitvarianten Belastungen.

In Kap. 6.1.2 wird eine Modifikation des hier untersuchten Verkehrsmodells vorgestellt, wobei die Leistungsfähigkeit einer Überlast-Abwehrstrategie ermittelt wird.

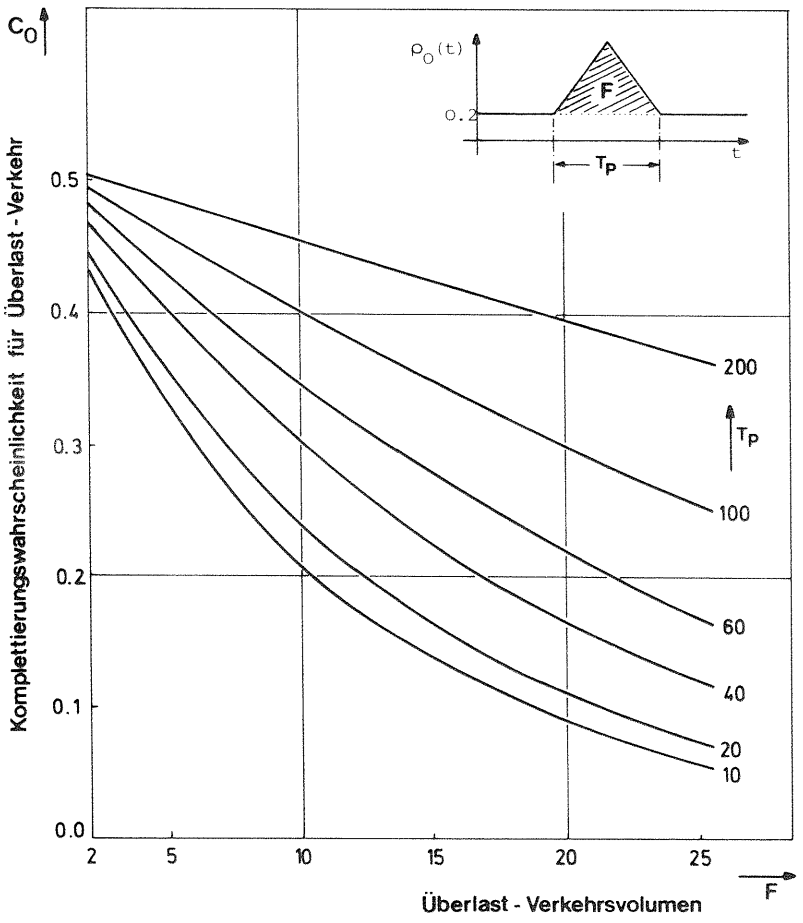


Bild 5.22 *Abhängigkeit der Systemleistung von der Dauer und vom Verkehrsvolumen der Überlastimpulse.*

Parameter : $N_O = 3$ $N_{LIM} = 4$
$N_1 = 30$ $\gamma = 0.1$

6. MODELLE FÜR ÜBERLASTABWEHRSTRATEGIEN

In diesem Kapitel werden Modelle zur Beschreibung von Überlastabwehrstrategien vorgestellt und untersucht. Die Modellbildung konzentriert sich hierbei auf zwei Klassen von Maßnahmen zur Überlastregelung:

- Die Drosselung der angebotenen Rufströme:
  Es handelt sich hier um Überlastabwehrstrategien, bei denen zur Vermeidung von Überlastentwicklungen vorsorglich Rufe abgewiesen werden. In diesem Zusammenhang werden zwei Grundmechanismen zur Rufabweisung untersucht: die Zweipunkt-Regelung der Rufaufnahme und die graduelle Rufblockierung.
- Die optimale Ausnutzung der vorhandenen Systemkapazität:
  Diese Kategorie von Überlastabwehrmethoden wird exemplarisch anhand einiger Modelle behandelt, bei denen die Leistung entwickelter Überlastabwehrmaßnahmen untersucht wird. Die Modellbildung umfaßt die Optimierung der dynamischen Speicherplatzreservierung und der effektiven Rechnerkapazität hinsichtlich der Rufkomplettierung.

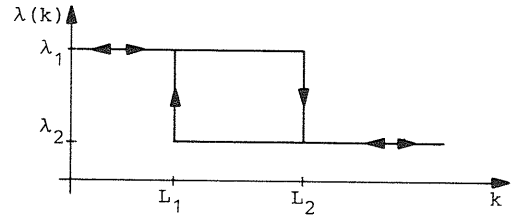
6.1 Drosselung der Rufannahme

6.1.1 Modell zur Rufannahmestrategie mit der Zweipunkt-Regelung

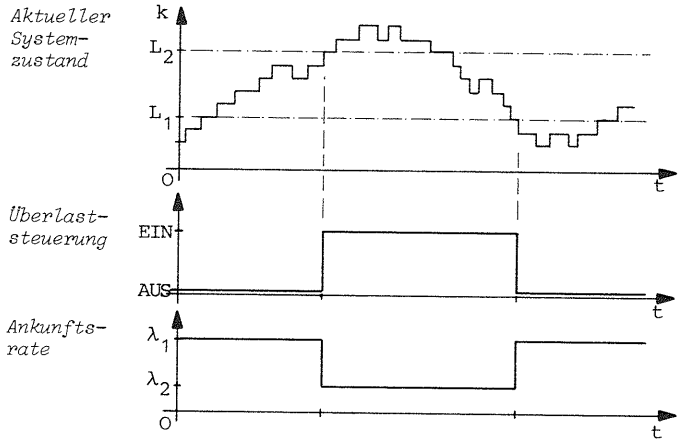
a) Modellbeschreibung

Die Drosselung der Rufe in einem Vermittlungssystem erfolgt i.a. mit einer differenzierten Ruf- bzw. Teilprozeßannahmestrategie. Dabei müssen der Einfluß des Rufwiederholungseffekts und der Verwaltungsaufwand zur Durchführung der Überlastabwehrmaßnahmen gering gehalten werden. Der Verwaltungsaufwand, der z.B. durch das Ein- und Ausschalten der Überlaststeuerung entsteht, kann sich bei einer ungünstigen Dimensionierung der Rufannahmestrategie erheblich erhöhen und auftretende Systemüberlastungen weiter verschärfen.

Im folgenden wird ein Modell vorgestellt, in dem eine Annahmestrategie für Anforderungen zur Teilprozeßaktivierung mittels einer Zweipunkt-Regelung untersucht wird. Der Mechanismus wird in Bild 6.1 dargestellt. Im normalen Betriebszustand des Vermittlungssystems sei λ_1 die Rate der Teilprozeßanforderungen, die zur Aktivierung von Software-Teilprozessen führen. Dazu gehören alle für die Rufverarbeitung erforderlichen Teilprozesse (vgl. Kap.2), wobei für die Überlaststeuerungsbetrachtung zwei Klassen von Anforderungen wesentlich sind: die Belegungsanforderungen - die zur



(a) Regelmechanismus



(b) Prozeßverlauf

Bild 6.1 *Drosselung des Teilprozeßverkehrs mit der Zweipunkt-Regelung.*

Aktivierung der Rufverarbeitungsteilprozesse führen - und die Anforderungen zur Rufauslösung. Die Anzahl k aller aktiven Teilprozesse im System wird als Überlastindikator genommen, da sie die Intensität der Steuerungsaufrufe und damit die Belastung der Steuerungseinheit bestimmt (vgl. Modellbildung in Kap. 6.2.2). Wird die Überlaststeuerung eingeschaltet, so erfolgt eine Drosselung des Teilprozeßverkehrs, wobei Teilprozesse, die für die Komplettierung bereits aktiver Rufe im System wichtig sind, vorrangig angenommen werden, z.B. Anforderungen zur Rufauslösung. Die gedrosselte Rate dieser Anforderungen sei λ_2 ($\lambda_2<\lambda_1$).

Bild 6.1a zeigt den Mechanismus der Überlastabwehrstrategie mittels der Zweipunkt-Regelung. Überschreitet die aktuelle Anzahl k aktiver Teilprozesse den oberen Schwellenwert L_2, so wird die Überlaststeuerung aktiviert und die Ankunftsrate der Teilprozeßanforderungen von λ_1 auf λ_2 gedrosselt. Die Systembelastung ist jetzt geringer, so daß das im System vorhandene Überlast-Verkehrsvolumen abgebaut werden kann. Werden bei der gedrosselten Rate λ_2 keine neuen Rufe mehr angenommen, so werden Folgeteilprozesse und Rufauslösungen vorrangig bearbeitet; eine Verringerung der Anzahl aktiver Rufe im System ist zu erwarten. Die Überlaststeuerung wird deaktiviert, wenn die Anzahl k aktiver Teilprozesse unter den Schwellenwert L_1 absinkt. Diese systemzustandsabhängige Überlaststeuerung wird in Bild 6.1b anhand des Prozeßverlaufs erläutert.

Das Verkehrsmodell ist in Bild 6.2, dessen Komponenten im folgenden näher betrachtet werden, schematisch dargestellt:

- Der normale Teilprozeßverkehr ist ein Poisson-Prozeß mit der Rate λ_1 (vgl. Gl. 5.6). Ist die Überlaststeuerung aktiv, so wird der Teilrufverkehr, der ebenfalls mit einem Poisson-Prozeß mit der Rate λ_2 modelliert wird, gedrosselt.
- Die maximale Anzahl der aktiven Teilprozesse im System ist auf n begrenzt. Unter Berücksichtigung aller Arten von Teilprozessen wird die Bedienungsdauer T_H (holding time), welche der aktiven Dauer eines Teilprozesses entspricht, als negativ exponentiell verteilte Zufallsvariable angenommen:

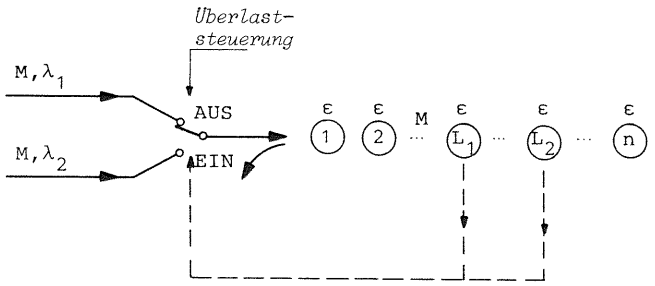


Bild 6.2 *Verkehrsmodell für die Drosselung des Teilprozeßverkehrs mit der Zweipunkt-Regelung.*

$$F_H(t) = P\{T_H \leq t\} = 1 - e^{-\varepsilon t},$$

$$E[T_H] = \frac{1}{\varepsilon}. \qquad (6.1)$$

Das System wird als Verlustsystem betrieben, d.h. falls alle Bedienungseinheiten belegt sind, werden ankommende Teilprozeßanforderungen abgewiesen. Dies beinhaltet eine weitere Begrenzung der Anzahl aktiver Teilprozesse im System, die einer Regelung der Teilrufverkehrsintensität entspricht.

b) Zustandsdiagramm und rekursive Lösung

Für die Modellanalyse werden folgende Symbole verwendet:

λ_1 normale Rate des Teilprozeßverkehrs.

λ_2 gedrosselte Rate des Teilprozeßverkehrs.

ε Enderate einer Bedienungseinheit für Teilprozesse.

n maximale Anzahl aktiver Teilprozesse im System.

$A_1 = \frac{\lambda_1}{\varepsilon}$ normales Verkehrsangebot der Teilprozesse.

$A_2 = \frac{\lambda_2}{\varepsilon}$ gedrosseltes Verkehrsangebot der Teilprozesse.

X Zufallsvariable für die aktuelle Anzahl aktiver Teilprozesse.

Z Hilfsvariable zur Kennzeichnung des Zustands der Überlaststeuerung:
Z = 1 deaktivierte Überlastabwehr
Z = 2 aktivierte Überlastabwehr.

P(i,j) = P{X=i,Z=j}
Zustandswahrscheinlichkeit dafür, daß i Teilprozesse aktiv sind und die Überlaststeuerung im Zustand j (j = 1,2) ist.

Der Zustandsprozeß des Systems läßt sich mit der Zufallsvariable X und der Hilfsvariable Z vollständig beschreiben. Da alle Modellkomponenten die Markoff-Eigenschaft aufweisen, kann für die Analyse im stationären Falle das Zustandsübergangsdiagramm in Bild 6.3 entwickelt werden, das im folgenden kurz erläutert wird:

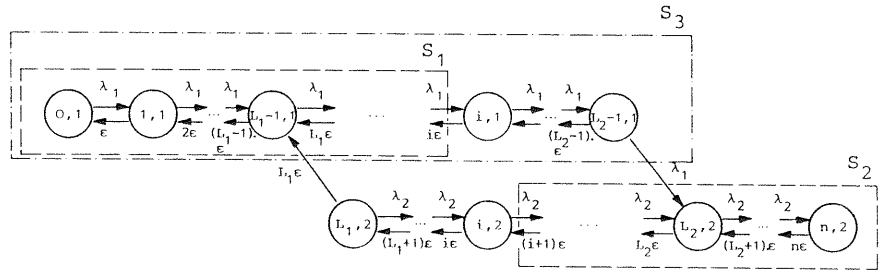


Bild 6.3 *Zustandsübergangsdiagramm des Modells für die Zweipunkt-Regelung des Teilprozeßverkehrs.*

- Befindet sich die Überlastabwehr im aktiven bzw. inaktiven Zustand, so ist die zugehörige Ankunftsrate λ_2 bzw. λ_1.
- Befinden sich X = i aktive Teilprozesse im System, so beträgt die Enderate $i\varepsilon$.
- Die Überlastabwehr kann im Zustand $(L_2-1,1)$ mit der Ankunft eines Teilprozesses (Rate λ_1) eingeschaltet und im Zustand $(L_1,2)$ mit der Endigung eines Teilprozesses (Rate $L_1\varepsilon$) abgeschaltet werden.
- Im Zustand $(n,2)$ werden ankommende Teilprozesse abgewiesen; sie führen daher zu keiner Zustandsänderung.

Betrachtet werden die Makrozustände S_1, S_2 und S_3, die in Bild 6.3 gekennzeichnet sind. Befindet sich der Makrozustand S_1 im statistischen Gleichgewicht, so erhält man gemäß Gl. (4.10) folgende Zustandsgleichungen:

$$\lambda_1 P(i-1,1) = i\varepsilon P(i,1), \quad i = 1,\ldots,L_1-1, \tag{6.2}$$

$$\lambda_1 P(i-1,1) = i\varepsilon P(i,1)+L_1\varepsilon P(L_1,2), \quad i = L_1,\ldots,L_2-1. \tag{6.3}$$

Analog lauten die Zustandsgleichungen für S_2

$$\lambda_2 P(i,2)+\lambda_1 P(L_2-1,1) = (i+1)\varepsilon P(i+1,2), \quad i=L_1,\ldots,L_2-1, \tag{6.4}$$

$$\lambda_2 P(i,2) = (i+1)\varepsilon P(i+1,2), \quad i = L_2,\ldots,n-1. \tag{6.5}$$

Für S_3 gilt:

$$L_1\varepsilon P(L_1,2) = \lambda_1 P(L_2-1,1). \tag{6.6}$$

Zusammen mit der Normierungsbedingung

$$\sum_{i=0}^{L_2-1} P(i,1) + \sum_{i=L_1}^{n} P(i,2) = 1 \tag{6.7}$$

bilden die Gleichungen (6.2-6.6) ein lineares Gleichungssystem zur Bestimmung der Zustandswahrscheinlichkeiten. Eine geschlossene Lösung kann angegeben werden; sie wird nachfolgend hergeleitet.

Man erhält durch sukzessives Einsetzen der Gleichungen (6.2) und (6.5)

$$P(i,1) = \frac{A_1}{i} P(i-1,1) = \frac{A_1^i}{i!} P(0,1), \quad i = 0,\ldots,L_1-1, \tag{6.8a}$$

$$P(i,2) = \frac{i+1}{A_2} P(i+1,2) = \frac{n!}{i!A_2^{n-i}} P(n,2), \quad i=L_2,\ldots,n \tag{6.8b}$$

oder speziell für die Zustandswahrscheinlichkeiten $P(L_1-1,1)$ und $P(L_2,2)$

$$P(L_1-1,1) = \frac{A_1^{L_1-1}}{(L_1-1)!} P(0,1), \tag{6.9a}$$

$$P(L_2,2) = \frac{n!}{L_2!A_2^{n-L_2}} P(n,2). \tag{6.9b}$$

Aus Gl. (6.3) ergibt sich ebenfalls durch sukzessives Einsetzen

$$P(i,1) = \frac{A_1}{i} P(i-1,1) - \frac{L_1}{i} P(L_1,2)$$

$$= \frac{A_1^{i-L_1+1}}{i!}(L_1-1)! P(L_1-1,1) - L_1 P(L_1,2) \sum_{j=0}^{i-L_1} \frac{A_1^j (i-j-1)!}{i!}.$$

Mit Gl. (6.9a) erhält man

$$P(i,1) = \frac{A_1^i}{i!} P(0,1) - L_1 P(L_1,2) \sum_{j=0}^{i-L_1} \frac{A_1^j (i-j-1)!}{i!}, \quad i=L_1,\ldots,L_2-1 \tag{6.10}$$

oder speziell für $i = L_2-1$

$$P(L_2-1,1) = P(0,1)\underbrace{\frac{A_1^{L_2-1}}{(L_2-1)!}}_{K_1} - P(L_1,2)\underbrace{\frac{L_1}{(L_2-1)!}\sum_{j=0}^{L_2-L_1-1} A_1^j (L_2-j-2)!}_{K_2} . \quad (6.11)$$

Analog erhält man mit Gl. (6.4) und (6.9b)

$$\begin{aligned} P(i,2) &= \frac{i+1}{A_2} P(i+1,2) - \frac{A_1}{A_2} P(L_2-1,1) \\ &= \frac{L_2!}{i!} \frac{1}{A_2^{L_2-1}} P(L_2,2) - \frac{A_1}{A_2} P(L_2-1,1) \sum_{j=0}^{L_2-i-1} \frac{(i+j)!}{i!} \frac{1}{A_2^j} \\ &= \frac{n!}{i!} \frac{1}{A_2^{n-i}} P(n,2) - \frac{A_1}{A_2} P(L_2-1,1) \sum_{j=0}^{L_2-i-1} \frac{(i+j)!}{i!} \frac{1}{A_2^j}, \end{aligned}$$

$$i = L_1, \ldots, L_2-1 \quad (6.12)$$

oder speziell für $i = L_1$

$$P(L_1,1) = P(n,2)\underbrace{\frac{n!}{L_1!}\frac{1}{A_2^{n-L_1}}}_{K_3} - P(L_2-1,1)\underbrace{\frac{A_1}{A_2}\sum_{j=0}^{L_2-L_1-1}\frac{(L_1+j)!}{L_1!}\frac{1}{A_2^j}}_{K_4} . \quad (6.13)$$

Aus Gl. (6.6) folgt:

$$P(L_1,2) = \underbrace{\frac{A_1}{L_1}}_{K_5} P(L_2-1,1). \quad (6.14)$$

Die Normierungsgleichung (6.7) läßt sich mit den Beziehungen (6.8a,b), (6.10) und (6.12) wie folgt umformen:

$$P(0,1)\underbrace{\sum_{i=0}^{L_2-1}\frac{A_1^i}{i!}}_{K_6} - P(L_1,2)\underbrace{L_1\sum_{i=L_1}^{L_2-1}\sum_{j=0}^{i-L_1}\frac{A_1^j(i-j-1)!}{i!}}_{K_7}$$

$$+ P(n,2)\underbrace{\sum_{i=L_1}^{n}\frac{n!}{i!A_2^{n-i}}}_{K_8} - P(L_2-1,1)\underbrace{\frac{A_1}{A_2}\sum_{i=L_1}^{L_2-1}\sum_{j=0}^{L_2-i-1}\frac{(i+j)!}{i!}\frac{1}{A_2^j}}_{K_9} = 1. \quad (6.15)$$

Die Gleichungen (6.11), (6.13), (6.14) und (6.15) stellen ein lineares Gleichungssystem zur Berechnung von $P(0,1)$, $P(L_2-1,1)$, $P(L_1,2)$ und $P(n,2)$ dar. Es ergibt sich :

$$\begin{cases} P(L_2-1,1) = \left[\frac{K_6(1+K_2K_5)}{K_1} - K_7K_5 + \frac{K_8(K_4+K_5)}{K_3} - K_9\right]^{-1} \\ P(0,1) = \frac{1+K_2K_5}{K_1} P(L_2-1,1) \\ P(L_1,2) = K_5 P(L_2-1,1) \\ P(n,2) = \frac{K_4+K_5}{K_3} P(L_2-1,1). \end{cases} \quad \ldots(6.16)$$

Anhand der Zustandswahrscheinlichkeiten in Gl. (6.16) lassen sich alle übrigen Zustandswahrscheinlichkeiten gemäß Gl. (6.8a,b), (6.10) und (6.12) explizit angeben. Die charakteristischen Größen werden aus den gewonnenen Zustandswahrscheinlichkeiten berechnet.

Durch die Aktivierung der Überlaststeuerung werden Belegungsanforderungen abgewiesen. Die Blockierungswahrscheinlichkeit B dieser Teilprozeßanforderungen ist die Summe aller Zustands-

wahrscheinlichkeiten, in denen sich die Überlaststeuerung im aktiven Zustand befindet:

$$B = \sum_{i=L_1}^{n} P(i,2). \qquad (6.17)$$

Die mittlere Anzahl aktiver Teilprozesse im System (mittlere Systembelegung) errechnet sich zu

$$E[X] = \sum_{i=1}^{L_2-1} i\, P(i,1) + \sum_{i=L_1}^{n} i\, P(i,2). \qquad (6.18)$$

c) Dimensionierungsbeispiel

Anhand des Bildes 6.4 wird ein Beispiel für die Dimensionierung der Schwellenwerte L_1 und L_2 der Zweipunkt-Regelung gezeigt. Die mittlere Systembelegung $E[X]$ und die Blockierungswahrscheinlichkeit B der Belegungsanforderungen werden für verschiedene Werte von L_1 und L_2 aufgetragen. Wie zu erwarten ist, führt eine Erhöhung der Schwellenwerte zu einer Reduzierung der Blockierung. Parallel steigt jedoch die mittlere Anzahl aktiver Teilprozesse und somit die Intensität des Teilrufverkehrs an. Durch eine geeignete Dimensionierung der Schwellenwerte können $E[X]$ und B optimiert werden.

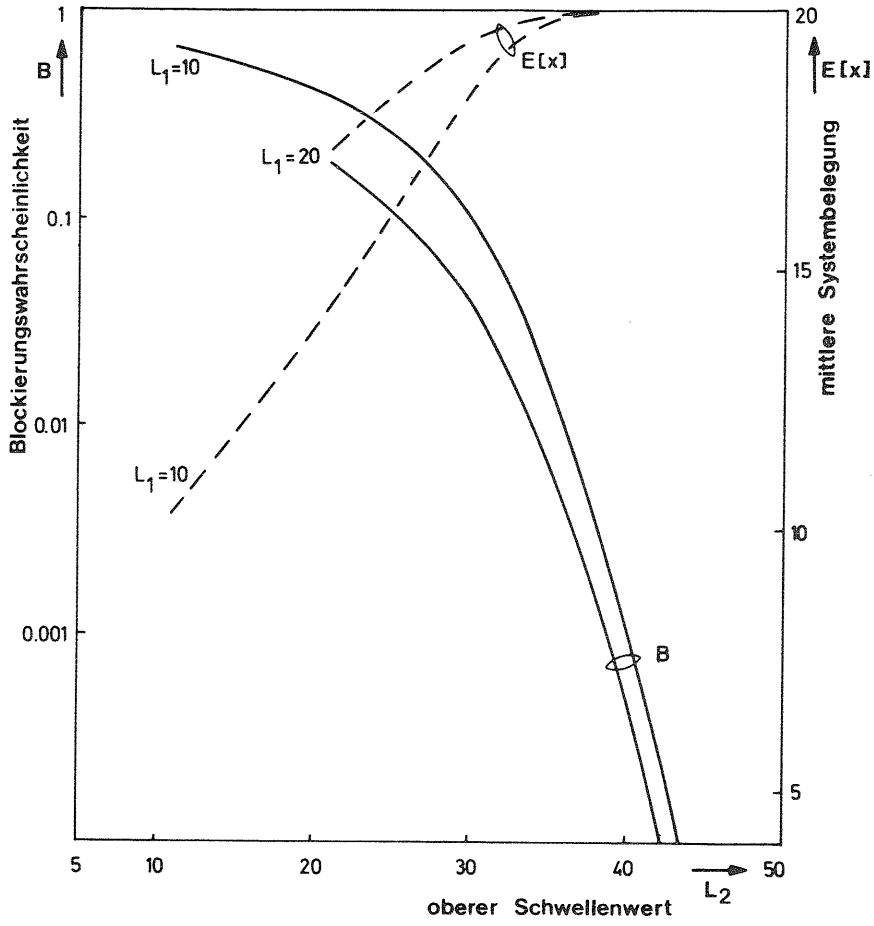


Bild 6.4 *Zur Dimensionierung der Schwellenwerte L_1 und L_2 für die Zweipunkt-Regelung.*
Parameter : $A_1 = 20$, $A_2 = 5$, $n = 50$

6.1.2 Graduelle Rufblockierung

a) Einführung einer Überlastabwehrstrategie im Modell mit wartezeitabhängiger Rufkomplettierung

Die Verringerung der Leistungsfähigkeit eines Vermittlungssystems bei einer Überlastung wurde in Kap. 5.2.2 diskutiert. Dort wird ein Verkehrsmodell vorgestellt, welches die Abhängigkeit der Rufkomplettierungswahrscheinlichkeit von der Wartezeit aufgrund des Teilnehmerverhaltens berücksichtigt. Die Systemantwort auf kurzzeitige, impulsförmige Überlastung wurde anhand dieses Modells untersucht. Der Grund für die Leistungssenkung liegt darin, daß die Überlastung eine Verlängerung der Wartezeiten verursacht, die zur häufigen Rufaufgabe bzw. Rufunterbrechung führt.

Im folgenden wird eine Überlastabwehrmaßnahme mit einer graduellen Rufblockierung vorgestellt, die durch eine Modifikation des Verkehrsmodells nach Kap. 5.2.2 untersucht wird. Dieses Verkehrsmodell ist vom Typ $M^{[X]}/M/1$ mit zustandsabhängiger Gruppenankunft. Die Modellkomponenten sind (vgl. Kap. 5.2.2b):

- Der Rufankunftsprozeß ist ein verallgemeinerter Poisson-Prozeß mit der Rate $\lambda(t)$.
- Ein Ruf, der zum Ankunftszeitpunkt k Phasen im System antrifft, wird mit der Wahrscheinlichkeit $g_j^{(k)}$ eine Anzahl j von Phasen zur Verarbeitung erzeugen. Diese Modellkomponente modelliert die wartezeitabhängige Rufverarbeitungszeit.
- Rufe mit j Phasen werden mit der Wahrscheinlichkeit c_j komplettiert.

Abhängig von der Anzahl k der Phasen im System, die ein Ruf zu seinem Ankunftszeitpunkt antrifft, wird nun eine zustandsabhängige Rufabweisungswahrscheinlichkeit B_k definiert, mit welcher der Ruf nicht angenommen wird. Ähnlich wie bei der Überlastabwehrstrategie nach dem Prinzip der Zweipunkt-Regelung in Kap. 6.1.1 werden zwei Schwellenwerte L_1 und L_2 eingeführt. Während für $k<L_1$ ankommende Rufe angenommen werden, werden alle Rufe bei

$k > L_2$ abgewiesen. In dieser Weise werden die Wartezeit eines Rufes bis zur Bearbeitung begrenzt und die angenommenen Rufe mit einer höheren Wahrscheinlichkeit komplettiert. Zwischen den Schwellenwerten L_1 und L_2 hat B_k einen ansteigenden Funktionsverlauf, der in dieser Untersuchung linear gewählt wird. Zusammengefaßt kann die graduelle Blockierung wie folgt formuliert werden (vgl. Bild 6.5):

$$B_k = \begin{cases} 0 & , k < L_1 \ , \\ \dfrac{k-L_1}{L_2-L_1} & , L_1 \le k \le L_2 \ , \\ 1 & , k > L_2 . \end{cases} \qquad (6.19)$$

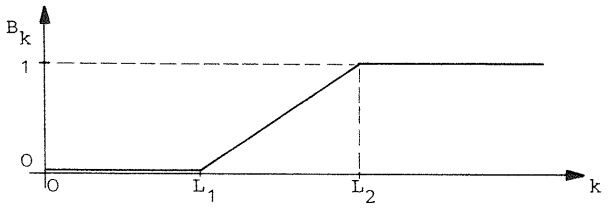


Bild 6.5 Graduelle Rufblockierung.

b) Modellmodifikation und -analyse

Aufgrund der Rufblockierungsstrategie nach Gl. (6.19) und der zustandsabhängigen Gruppenverteilung $g_j^{(k)}$ nach Gl. (5.51) hat die Warteschlange des Verkehrsmodells eine begrenzte Kapazität. Die maximal mögliche Anzahl der Phasen im System entsteht, wenn ein Ruf, der angenommen wird, L_2-1 Phasen antrifft und die Anzahl N_1 von Phasen zur Verarbeitung generiert. Die Systemkapazität ist damit L_2+N_1-1 und die Warteschlangenlänge beträgt $S_0=L_2+N_1-2$. Dies entspricht einem modifizierten Modell des Typs $M^{[X]}/M/1\text{-}S_0$ mit zustandsabhängigem Gruppenankunftsprozeß.

Mittels einer Modifikation der nach Gl. (5.51) definierten Ver-

teilung der Anzahl von Phasen $g_j^{(k)}$ kann die in Kap. 5.2.2d dargestellte Modellanalyse auch unter Einbeziehung der hier vorgestellten Überlaststeuerung angewendet werden [70]. Dabei wird die graduelle Rufblockierungsstrategie in der modifizierten Gruppenverteilung $\bar{g}_j^{(k)}$ durch die Festlegung berücksichtigt, daß abgewiesene Rufe keine Phase generieren (j=0). Man erhält den folgenden Zusammenhang zwischen $\bar{g}_j^{(k)}$ und $g_j^{(k)}$:

$$\begin{cases} \bar{g}_0^{(k)} = B_k & , j = 0 , \\ \bar{g}_j^{(k)} = g_j^{(k)} \cdot (1-B_k) & , j > 0 . \end{cases} \tag{6.20}$$

Die bedingte Komplettierungswahrscheinlichkeit für abgewiesene Rufe ist

$$c_0 = 0. \tag{6.21}$$

Da hier die Anzahl der Zustände (vgl. Zustandsübergangsdiagramm in Bild 5.19) begrenzt ist, vereinfacht sich die numerische Auflösung des Systems der Zustandsgleichungen gemäß Gl. (5.54) bzw. (5.59) sowohl für die stationäre als auch für die instationäre Analyse.

Aus den gewonnenen zeitabhängigen Zustandswahrscheinlichkeiten $P_k(t)$, k = 0, 1,...,S+1, werden - ähnlich wie in den Gleichungen (5.55-5.58) - die Systemcharakteristiken berechnet. Die mittlere Anzahl der Phasen im System lautet:

$$E[X(t)] = \sum_{k=1}^{S_0+1} kP_k(t). \tag{6.22}$$

Mit der zeitabhängigen Rufkomplettierungswahrscheinlichkeit

$$C(t) = \sum_{k=0}^{S_0+1} P_k(t) \sum_{j=N_0}^{N_1} c_j \bar{g}_j^{(k)} \tag{6.23}$$

kann die transiente Rufkomplettierungsrate angegeben werden

$$Y(t) = \lambda(t) \cdot C(t). \tag{6.24}$$

Die normierte Rufkomplettierungsrate ist

$$Y_0(t) = Y(t).h.E[G^{(k_0)}]. \tag{6.25}$$

c) Instationäre Systemantwort und Leistung der Überlastabwehrstrategie

Die Leistungsbeurteilung der Überlastabwehrstrategie nach dem Prinzip der graduellen Rufabweisung (Gl. 6.19) wird nachfolgend anhand der Systemantwort auf dreieckförmige Überlastmuster durchgeführt. Bild 6.6 zeigt einen Vergleich der dynamischen Systemantworten mit und ohne Überlastabwehr, wobei die Trajektorien der zeitabhängigen Rufkomplettierungsrate dargestellt sind. In diesem Diagramm wird ebenfalls die Rufkomplettierungscharakteristik im stationären Falle gezeigt. Bei der stationären Betrachtung ist für höhere Belastungen eine Erhöhung der Rufkomplettierungsrate zu erkennen. Die Ein- und Ausschwingphasen der transienten Systemantwort werden in Bild 6.7 verglichen. Es ist ersichtlich, daß die Überlastabwehrstrategie einen flacheren Anstieg der Rufkomplettierungsrate in der Einschwingphase und eine deutliche Leistungsverbesserung in der Ausschwingphase bewirkt. Durch die Drosselung der angebotenen Belastung wird die Überlastabbauphase kürzer.

Mit dem in den Gleichungen (5.62-5.65) definierten Gütefaktor C_0 kann ein Leistungsvergleich hinsichtlich der gesamten instationären Systemreaktion durchgeführt werden. Bild 6.8 zeigt die Komplettierungswahrscheinlichkeit des Überlastverkehrs als Funktion des Überlast-Verkehrsvolumens F mit der Überlastdauer T_p als Parameter. Eine Leistungssteigerung wird durch die Überlastabwehrstrategie bei starken Überlastungen erreicht (F>10). Für kleinere Werte des Überlast-Verkehrsvolumens, d.h. bei relativ schwachen Überlastungen, bewirkt die Überlastabwehr jedoch eine geringfügige Leistungssenkung. Der Grund liegt darin, daß hier durch eine zu früh einsetzende Überlastabwehr der Ankunftsverkehr bereits gedrosselt wird, obwohl noch keine Überlastsituation vorliegt.

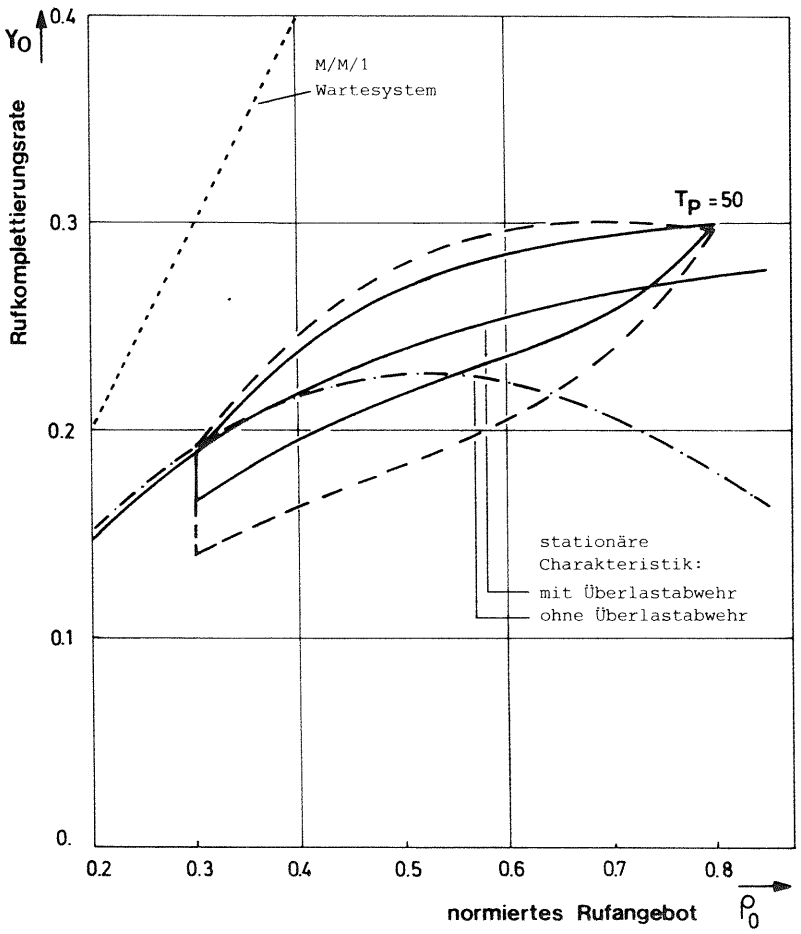


Bild 6.6 *Trajektorien der transienten Systemantwort auf einen dreieckförmigen Überlastimpuls.*

——— *ohne Überlastabwehr*
——— *mit Überlastabwehr*

Parameter : $N_O = 3$ $N_{LIM} = 4$ $L_1 = 25$
$N_1 = 30$ $\gamma = 0.1$ $L_2 = 35$

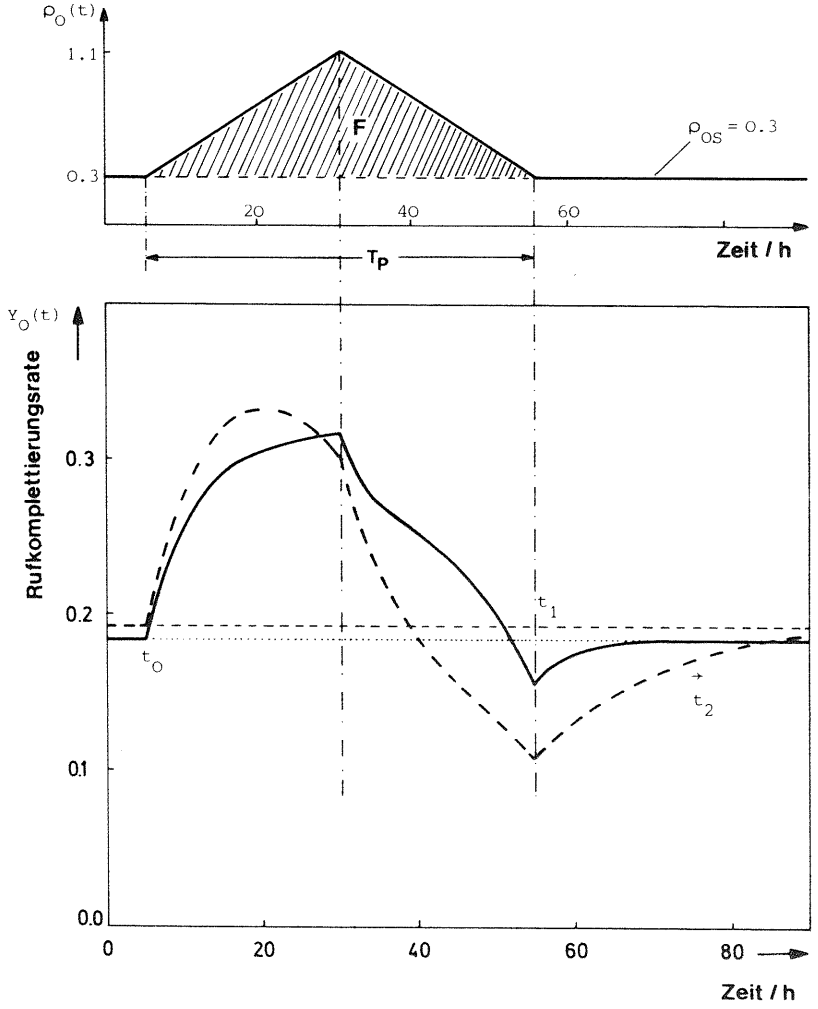


Bild 6.7 *Systemantwort auf einen dreieckförmigen Überlastimpuls.*

——— *ohne Überlastabwehr*
——— *mit Überlastabwehr*

Parameter : $N_O = 3$ $N_{LIM} = 4$ $L_1 = 25$
$N_1 = 30$ $\gamma = 0.1$ $L_2 = 35$

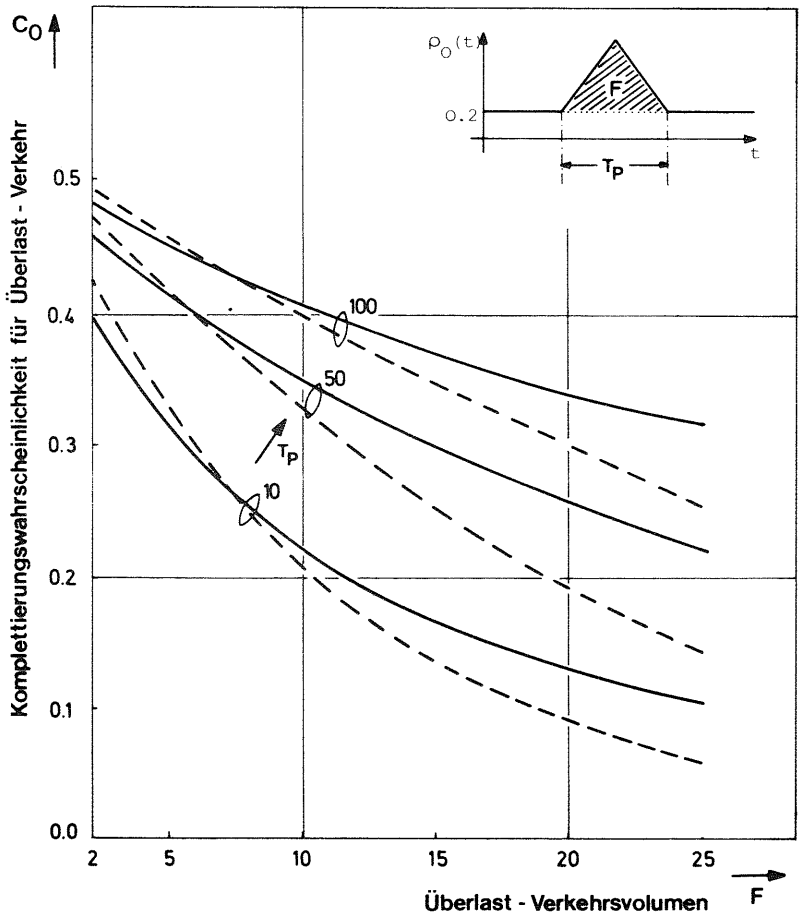


Bild 6.8 *Zur Leistungsbeurteilung der Überlastabwehrstrategie.*

——— *ohne Überlastabwehr*

——— *mit Überlastabwehr*

Parameter : $N_O = 3$ $N_{LIM} = 4$ $L_1 = 25$
$N_1 = 30$ $\gamma = 0.1$ $L_2 = 35$

6.2 Optimale Ausnutzung der Systemkapazität

6.2.1 Optimierung der Teilprozeß-Aktivierung

a) Serielle Teilprozeß-Aktivierung in Software-Maschinen

In Vermittlungssystemen mit modularer Software-Struktur kann ein aktiver Ruf parallel und seriell mehrere Teilprozesse aktivieren, für welche rufbezogene Datenblöcke (CCBs: call control blocks) in Software-Maschinen (SMn) bereitgestellt werden (vgl. Kapitel 2.3.3). Der kritischere Fall ist dabei die serielle Aktivierung von Teilprozessen. Die hier auftretende sekundäre Rufblockierung kann die Leistungsfähigkeit des Systems ungünstig beeinflussen und muß deshalb bei der Dimensionierung von SMn berücksichtigt werden.

Im folgenden wird ein Grundmodell untersucht, in dem die serielle Rufverarbeitung in zwei Software-Maschinen abgebildet wird (Bild 6.9). Ein Ruf wird hier zunächst von SM_1 bearbeitet (z.B. für Verbindungsaufbau), falls in dieser SM ein freier CCB zur Verfügung steht. Wenn alle CCBs in SM_1 besetzt sind, wird der einge-

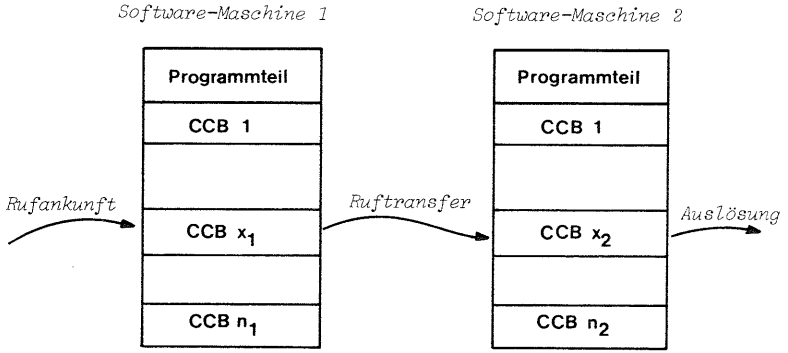


Bild 6.9 *Rufannahme und Teilprozeß-Aktivierung in zwei Software-Maschinen.*

troffene Ruf abgewiesen (primäre Blockierung). Nach der Verarbeitung in SM_1 wird der Ruf zur SM_2 (z.B. für Verbindungsüberwachung) transferiert. Sind alle CCBs in dieser SM zum Transferzeitpunkt besetzt, so tritt eine sekundäre Rufblockierung auf. In diesem Falle ist die Verarbeitungszeit dieses Rufes in SM_1 ineffektiv, da der Ruf nicht komplettiert werden kann. In Überlastsituationen, in denen die meisten CCBs belegt sind, kann dieser Effekt eine erhebliche Verringerung der Systemleistung verursachen.

b) Modell der Teilprozeß-Aktivierung zweier SMn

Das in Bild 6.10 dargestellte Verkehrsmodell umfaßt zwei aus Bedienungseinheiten zusammengesetzte Stufen, welche die CCBs in den SMn darstellen. Die Größen n_1 und n_2 sind dabei die zu dimensionierende Anzahl von CCBs in SM_1 und SM_2. Unter Berücksichtigung aller Ruftypen (normale Verbindung, Verbindung mit speziellen Leistungsmerkmalen, usw.) werden die Bedienungsdauern in SM_1 (Rufaufbauphase) und SM_2 (Gesprächsphase) als Zufallsvariablen mit negativ exponentiellen Verteilungen angenommen. Ein von SM_1 akzeptierter Ruf wird nach der Verarbeitung in dieser SM mit der Fortsetzungswahrscheinlichkeit Θ

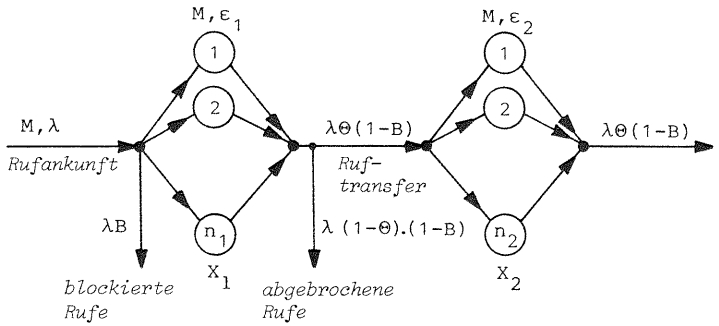


Bild 6.10 *Modell der seriellen Teilprozeß-Aktivierung in zwei Software-Maschinen.*

in die zweite Phase eintreten und übernommen werden. Rufe mit der komplementären Wahrscheinlichkeit $(1-\Theta)$ stellen abgebrochene Verbindungen dar, z.B. nach unvollständigem oder fehlerhaftem Wahlvorgang oder falls Tln. B besetzt ist u.ä.m..

Der aktuelle Belegungszustand von SM_1 bzw. SM_2 wird mit der Zufallsvariable X_1 bzw. X_2 beschrieben. Der ankommende Rufstrom wird durch einen Poisson-Prozeß mit der Rate λ modelliert.

Zur Vermeidung der diskutierten sekundären Rufblockierung wird eine Rufannahme-Strategie entwickelt, die einer Vorreservierung von CCBs in der zweiten SM entspricht [47]. Ein Ruf wird demnach angenommen, wenn gleichzeitig die folgenden Bedingungen zum Ankunftszeitpunkt erfüllt sind :

1) $x_1 < n_1$ (6.26a)
   Begrenzung wegen der Kapazität der SM_1.

2) $x_1 + x_2 < n_2$ (6.26b)
   Begrenzung der Anzahl von Rufen in den beiden SMn.

Dabei sind x_1 und x_2 Realisierungen der Zufallsvariablen X_1 und X_2 im statistischen Sinne.

Die in Gl.(6.26 a,b) formulierte Rufannahme-Strategie folgt dem Prinzip, die vorhandene Systemkapazität vorrangig den bereits angenommenen und noch komplettierbaren Rufen bereitzustellen, bevor neue Rufe angenommen werden.

c) Modellanalyse

Bei der Analyse werden folgende Symbole für die Modellkenngrößen angewendet :

λ Rufankunftsrate
ε_i Enderate der Bedienungseinheiten in SM_i
n_i Anzahl von CCBs in SM_i
$A_2 = \frac{\lambda}{\varepsilon_2}$ normiertes Verkehrsangebot für SM_2

Θ Wahrscheinlichkeit für Ruffortsetzung.

X_i Zufallsvariable für die aktuelle Anzahl belegter CCBs in SM_i.

$P(x_1,x_2) = P\{X_1=x_1,\ X_2=x_2\}$

Zustandswahrscheinlichkeit für x_1 belegte CCBs in SM_1 und x_2 belegte CCBs in SM_2.

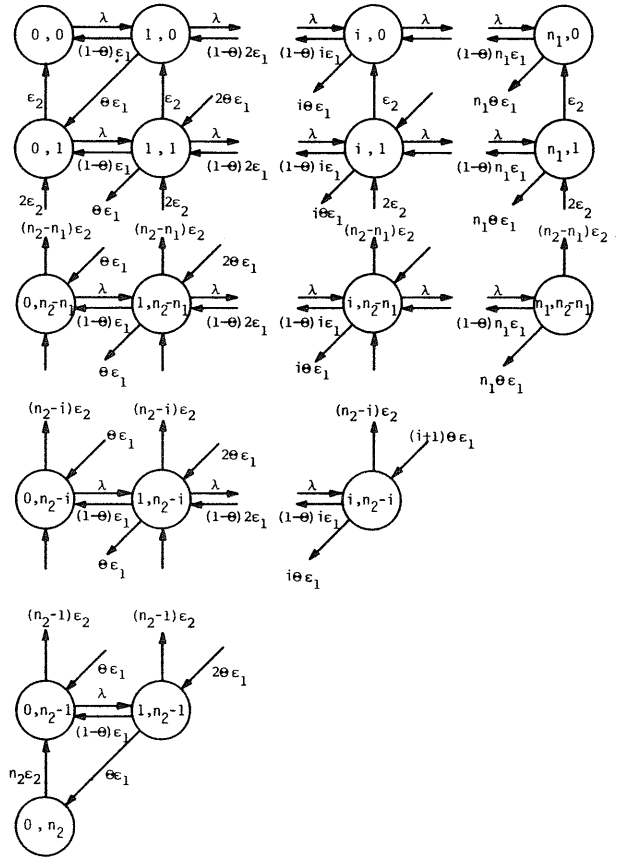


Bild 6.11 *Zustandsübergangsdiagramm.*

Das Verkehrsgeschehen und der Systemzustandsprozeß in den SMn können durch die ZV X_1 und X_2 bzw. den Zustand $(X_1=x_1,X_2=x_2)$ vollständig beschrieben werden. Da alle Ankunfts- und Bedienungsprozesse die Markoff-Eigenschaft (vgl.Gl.4.2) besitzen, kann der Zustandsprozeß mit dem Zustandsübergangsdiagramm eines zweidimensionalen Markoff-Prozesses beschrieben werden. Das in Bild 6.11 dargestellte Zustandsübergangsdiagramm wird exemplarisch durch nähere Betrachtung des Zustands (i,n_2-n_1) erläutert. In diesem Zustand können folgende Ereignisse zu einer Zustandsänderung führen :

- Ankunft eines Rufes : das System geht mit der Übergangswahrscheinlichkeitsdichte (ÜWD) λ in den Zielzustand $(i+1,n_2-n_1)$ über.
- Auslösung eines Rufes (SM_2) : mit der ÜWD $(n_2-n_1)\varepsilon_2$ wird der Zielzustand (i,n_2-n_1-1) erreicht.
- Endigung eines Rufes in SM_1 :
  - mit der ÜWD $\Theta i\varepsilon_1$ geht das System in den Zustand $(i-1,\ n_2-n_1+1)$ über (Rufübergabe).
  - mit der ÜWD $(1-\Theta)i\varepsilon_1$ geht das System in den Zustand $(i-1,n_2-n_1)$ über (Rufaufgabe).

Unter stationären Bedingungen befinden sich die Zustände im statistischen Gleichgewicht, wodurch gemäß Gl.(4.10) die folgenden Zustandsgleichungen gewonnen werden :

$$P(x_1,x_2).(\lambda+\varepsilon_1x_1+\varepsilon_2x_2) = \lambda P(x_1-1,x_2) + (1-\Theta)(x_1+1)\varepsilon_1P(x_1+1,x_2) + \Theta(x_1+1)\varepsilon_1P(x_1+1,x_2-1) + (x_2+1)\varepsilon_2\,P(x_1,x_2+1),$$

$$x_1 < n_1,\ x_1+x_2 < n_2, \qquad (6.27a)$$

$$P(n_1,x_2).(\varepsilon_1n_1+\varepsilon_2x_2) = \lambda P(n_1-1,x_2) + (x_2+1)\varepsilon_2P(n_1,x_2+1),$$

$$x_2 < n_2-n_1, \qquad (6.27b)$$

$$P(x_1,x_2).(\varepsilon_1 x_1+\varepsilon_2 x_2) = \lambda P(x_1-1,x_2) + \Theta(x_1+1)\varepsilon_1 P(x_1+1,x_2-1)\,,$$

$$x_2 \leq n_1,\ x_1+x_2=n_2, \qquad (6.27c)$$

wobei $P(x_1,-1) = P(-1,x_2) = P(n_1+1,x_2) = 0$ für alle x_1,x_2.

Zusammen mit der Normierungsbedingung gemäß Gl.(4.8b)

$$\sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2-x_1} P(x_1,x_2) = 1 \qquad (6.28)$$

stellen die Gleichungen (6.27a-c) ein lineares Gleichungssystem dar, das zur Bestimmung der Zustandswahrscheinlichkeiten $P(x_1,x_2)$ benutzt wird. Dabei wird eine numerische Iterationsmethode mit Überrelaxation angewendet [75].

Die primäre Blockierungswahrscheinlichkeit der Rufe lautet :

$$B = \sum_{x_2=0}^{n_2-n_1-1} P(n_1,x_2) + \sum_{x_1=0}^{n_1} P(x_1,n_2-x_1)\ . \qquad (6.29)$$

Die zwei Terme von B entsprechen den Rufannahme-Bedingungen nach Gl.(6.26a,b).

d) Ergebnisse und Anwendung für Dimensionierungsprobleme

Anhand numerisch gewonnener Ergebnisse wird der Einfluß von Systemparametern auf die Rufblockierung untersucht mit dem Ziel, eine optimale Dimensionierung von n_1 und n_2 für verschiedene Lastsituationen zu erreichen. Die Belastung ist hier normiert durch $\frac{A}{n_2}2=\frac{\lambda}{\varepsilon_2 n_2}$.

Bild 6.12 zeigt die Rufblockierung als Funktion der Anzahl n_1 von CCBs in SM_1, wenn n_2 konstant bleibt. Es ist ersichtlich, daß sich durch die Erhöhung von n_1 die Blockierungswahrscheinlichkeit nicht beliebig verringern läßt. Der Effekt kommt daher, daß oberhalb

eines - fast lastunabhängigen - Wertes für n_1 Rufe vorwiegend durch eine zweite Rufannahme-Bedingung nach Gl. (6.26b) blockiert werden. Dies wird auch in Bild 6.13 gezeigt, in dem der Einfluß der Fortsetzungswahrscheinlichkeit Θ untersucht wird.

In Bild 6.14 wird der Fall diskutiert, daß eine vorgegebene Anzahl n von CCBs auf zwei SMn verkehrsgerecht aufgeteilt werden soll ($n = n_1+n_2$). Die Blockierungswahrscheinlichkeit besitzt ein Minimum, das einer optimalen Dimensionierung für n_1 und n_2 entspricht. Das Minimum für B ist weniger empfindlich auf Belastungsänderungen als auf Schwankungen von Θ, n_1 bzw. n_2.

Eine Modifikation der Rufannahme-Strategie, die zu einer weiteren Reduzierung der Rufblockierungswahrscheinlichkeit im Überlastfall führt, wird in [47] vorgeschlagen und untersucht. Nach der modifizierten Strategie werden Rufe angenommen, auch wenn eine Vorreservierung eines CCB in SM_2 augenblicklich nicht möglich ist. Nach dem Ruftransfer wird der CCB in SM_1 dem Ruf weiterhin für die Verarbeitung in SM_2 zugeordnet, wobei das Prinzip der totalen Datentrennung in SMn kurzzeitig verletzt wird. Diese modifizierte Strategie ist anwendbar für den Fall, in dem die betrachteten SMn in derselben Steuerungseinheit lokalisiert sind.

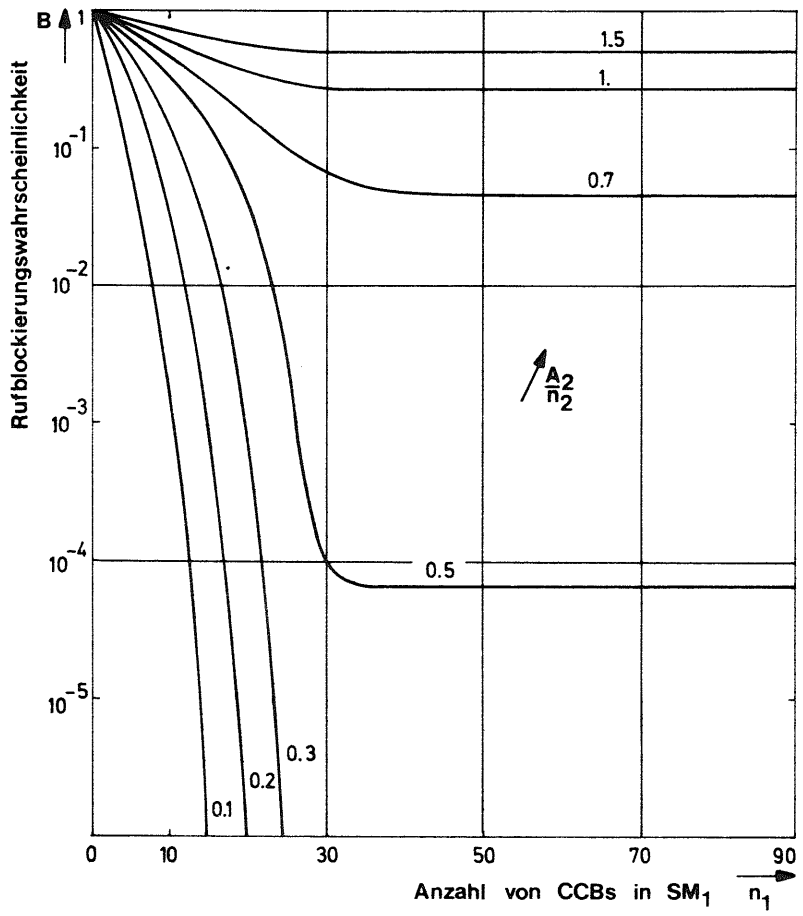


Bild 6.12 *Primäre Rufblockierungswahrscheinlichkeit.*

Parameter : $\frac{\varepsilon_1}{\varepsilon_2} = 3$, $n_2 = 90$

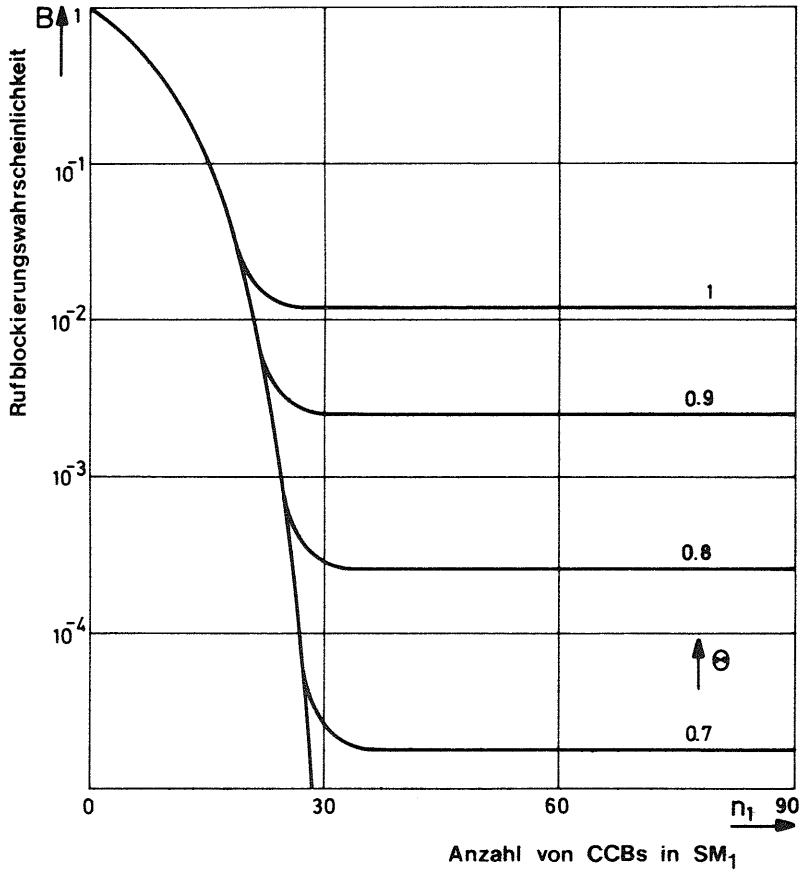


Bild 6.13 *Einfluß der Ruffortsetzungswahrscheinlichkeit.*

Parameter : $\frac{A_2}{n_2} = 0.7$, $\frac{\varepsilon_1}{\varepsilon_2} = 5$, $n_2 = 90$

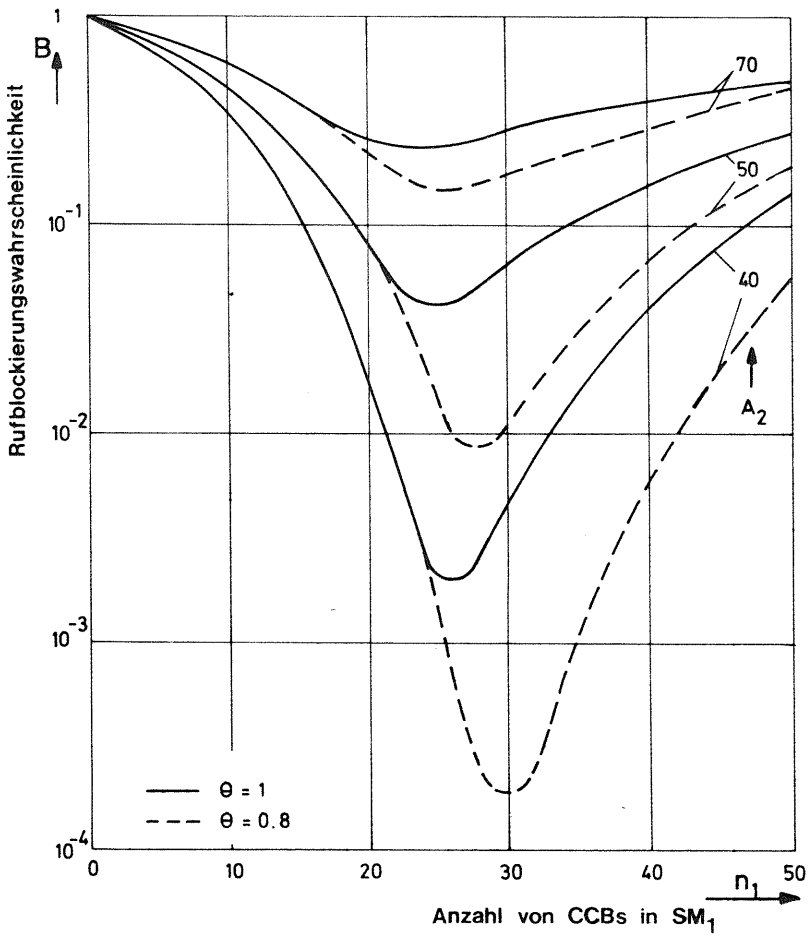


Bild 6.14 *Zur optimalen Aufteilung von CCBs in SMn.*
Parameter : $n_1 + n_2 = 100$

$$\frac{\varepsilon_1}{\varepsilon_2} = 3$$

6.2.2 Optimierung der Teilrufverarbeitung durch Zwangsauslösung nicht-komplettierbarer Rufe

a) Gesamtmodell für Ruf- und Teilrufverarbeitung

Die Rufkomplettierungscharakteristik ist ein wesentliches Merkmal für die Leistungsbeurteilung rechnergesteuerter Fernsprechvermittlungssysteme. Die Komplettierung eines Rufes hängt in erster Linie davon ab, wie seine Teilrufe verarbeitet werden. Da in modular strukturierten Vermittlungssystemen die Anzahl der Steuerungsaufrufe bzw. Teilrufe pro Ruf im Vergleich zu herkömmlichen Systemen höher ist, ist die Güte der Teilrufverarbeitung maßgeblich für die Leistungsfähigkeit eines Vermittlungssystems, insbesondere in Überlastsituationen.

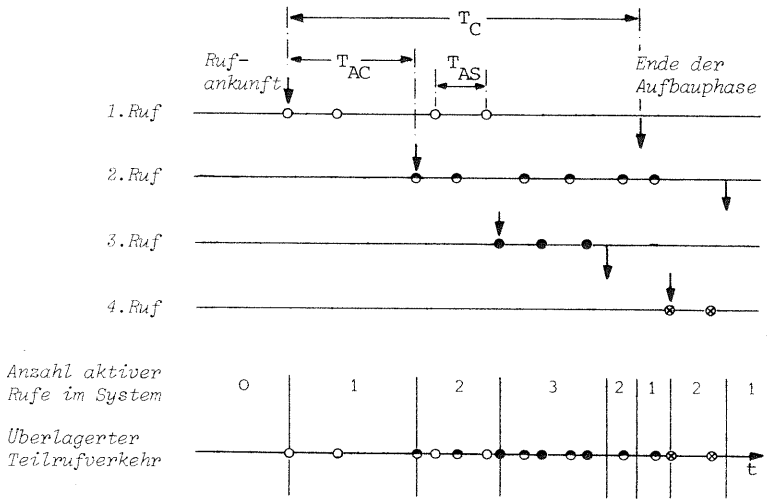


Bild 6.15 *Teilruferzeugung und überlagerter Teilrufverkehr (Rufaufbauphase).*

Teilrufe werden in einigen Modellansätzen und Untersuchungen berücksichtigt [34-38]. Die Erzeugung von Teilrufen wird in [34] mit Hilfe von Teilnehmer-Ereignisketten ausführlich diskutiert. In [35] und [36] werden simulative Untersuchungen über die Verarbeitung von Teilbelegungen (Teilrufen) in Vermittlungssystemen mit zentraler Steuerungsstruktur durchgeführt. Unter Berücksichtigung von Teilrufen und Steuerungsaufrufen werden in [37] und [38] grundlegende Mechanismen zur Überlastregelung vorgestellt.

Die Wechselwirkung zwischen Ruf- und Teilrufverarbeitung wird hier in einem Gesamtmodell beschrieben, mit dem die Leistung einer Überlastabwehrmaßnahme untersucht wird.

b) Modellbeschreibung und Parameterfestlegung

Bild 6.15 zeigt schematisch die Teilruferzeugung und den Teilrufstrom, welcher den Eingangsprozeß für den Prozessor darstellt.

Zur Kennzeichnung des Verkehrsgeschehens auf der Rufebene werden folgende Zufallsvariablen (ZVn) definiert :

- T_{AC} : Zwischenankunftsabstand für Rufe (T_{AC} : interarrival time for calls). Der Rufstrom wird hier durch einen Poisson-Prozeß mit der Rate λ modelliert :

$$F_{AC}(t) = P\{T_{AC} \leq t\} = 1 - e^{-\lambda t} ,$$

$$E[T_{AC}] = \frac{1}{\lambda} . \qquad \ldots(6.30)$$

- T_C : Dauer der Rufaufbauphase. Für die teilruforientierte Modellbildung genügt es, die Rufverarbeitung in der Aufbauphase zu betrachten, da in dieser Phase die meisten Teilrufe erzeugt werden. Betrachtet man nun alle Ruftypen (interne und externe Verbindung, Kurzwahl..), so kann T_C approximativ mit einer negativ exponentiellen Verteilung beschrieben werden :

$$F_C(t) = P\{T_C \leq t\} = 1 - e^{-\varepsilon t} ,$$

$$E[T_C] = \frac{1}{\varepsilon} . \qquad \ldots(6.31)$$

Auf der Teilrufebene wird folgende ZV eingeführt :

- T_{AS} : Zwischenankunftsabstand der Teilrufe, die zu einem Ruf gehören (T_{AS}: interarrival time for subcalls). Obwohl in der Realität die Zwischenankunftsabstände der Teilrufe voneinander abhängig sind (vgl. Kap. 5.1.2), ist die Beschreibung mittels einer unabhängigen ZV eine gute Näherung, wenn die mittlere Anzahl $E[G]$ der Teilrufe pro Ruf genügend groß ist. Betrachtet man die hohe Anzahl der Meldungen bzw. Steuerungsaufrufe, die in modular strukturierten Vermittlungssystemen pro Ruf erzeugt werden, so kann T_{AS} ebenfalls mit einer negativ exponentiellen Verteilung approximiert werden :

$$F_{AS}(t) = P\{T_{AS} \leq t\} = 1 - e^{-\lambda_S t} ,$$

$$E[T_{AS}] = \frac{1}{\lambda_S} . \qquad \ldots(6.32)$$

Teilrufe, die von den aktiven Rufen erzeugt werden, überlagern sich und bilden den Teilrufstrom. Der Teilrufprozeß ist somit abhängig von dem Rufprozeß, da die Rate des Teilrufstroms von der Anzahl der aktiven Rufe im System abhängt.

Das Gesamtmodell für Ruf- und Teilrufverkehr ist in Bild 6.16 dargestellt. Ein angenommener Ruf aktiviert eine Teilrufquelle, die Teilrufe mit der Rate λ_S produziert. Die Anzahl der Teilrufquellen im System ist begrenzt auf n. Sie entspricht der maximalen Anzahl aktiver Rufe im System. Nach der Rufaufbauphase (T_C) wird die Teilrufquelle abgeschaltet und freigegeben.

Die aktiven Teilrufquellen, deren Anzahl mit der ZV X_C beschrieben wird, erzeugen Teilrufe, die den Teilrufverkehr formen (vgl. Bild 6.16). Da die einzelnen aktiven Teilrufquellen nach Gl.(6.32) Poisson-Prozesse mit der Rate λ_S darstellen, ist der durch Superposition entstehende Teilrufverkehr (Bild 6.15) während der Zeitspanne, in der X_C konstant bleibt, ein Poisson-Prozeß mit der Rate $X_C \lambda_S$.

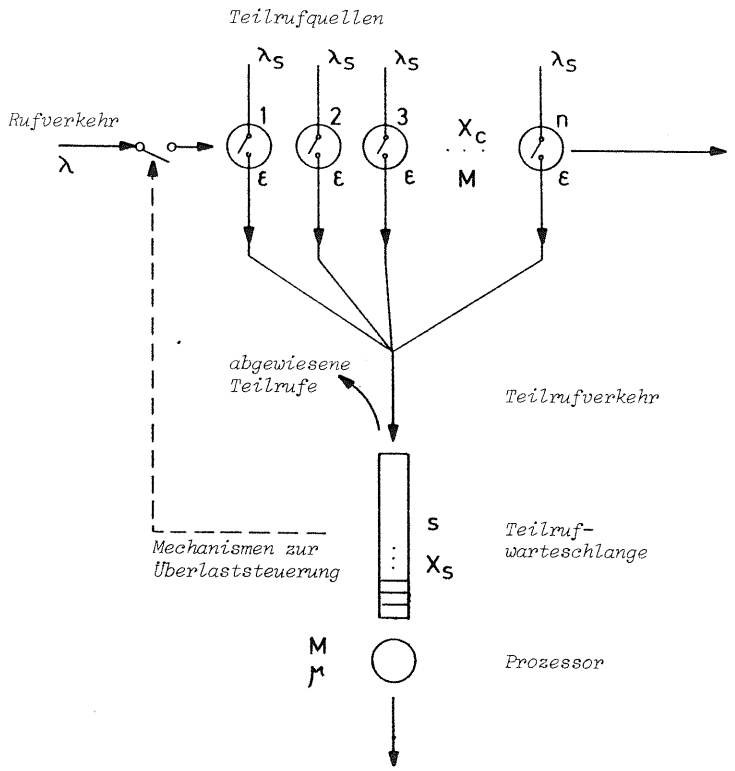


Bild 6.16 *Gesamtmodell für Ruf- und Teilrufverarbeitung.*

Der Teilrufverkehr bildet den Eingangsprozeß des einstufigen Warteschlangensystems (Bild 6.16), bestehend aus einer Bedienungseinheit, welche den Rechner bzw. die Steuerungseinheit modelliert, und einer Warteschlange mit S Warteplätzen, die den Pufferspeicher für vermittlungstechnische Ereignisse repräsentiert. Die Bedienungsdauer für Teilrufe wird als ZV mit negativ exponentieller Verteilung (Rate μ) angenommen.

Findet ein Teilruf das System in belegtem Zustand, d.h. sind der Prozessor und der Pufferspeicher besetzt, so wird er abgewiesen. Die Blockierung der Teilrufe hat folgende Auswirkungen :

- Ein Ruf, der einen blockierten Teilruf erzeugt hat, kann vom System nicht mehr komplettiert werden. Die Verarbeitungszeit derjenigen Teilrufe, die zu diesem Ruf gehören, trägt zur ineffektiven Ausnutzung der Prozessorzeit bei. Dieser Effekt führt zu einer Senkung der Rufkomplettierungsrate im Vermittlungssystem.
- Da die Signalisierung auf der peripheren Ebene langsamer als die Teilrufverarbeitung abgewickelt wird, kann ein nichtkomplettierbarer Ruf weitere Teilrufe produzieren, die nicht mehr zur Rufkomplettierung führen können. Dies hat eine weitere erhebliche Senkung der Systemleistung zur Folge, die in Überlastsituationen besonders kritisch wird.

c) Zwangsauslösung nichtkomplettierbarer Rufe

Eine einfache Maßnahme zur Erhöhung der Rufkomplettierung besteht darin, weitere Teilrufe von einem Ruf, der einen blockierten Teilruf erzeugt hat, nicht mehr anzunehmen. Dies geschieht z.B., indem der zugehörige Teilprozeß unmittelbar nach der Teilrufblockierung in den inaktiven Zustand versetzt wird. Die Realisierung dieser Maßnahme kann auch bei der Abtastung und Umsetzung vermittlungstechnischer Signale an der peripheren Schnittstelle erfolgen, indem z.B. durch eine Änderung der Maskierung die Abtastinformationen des betreffenden Rufes unterdrückt werden.

Für die Ausführung der Zwangsauslösung nichtkomplettierbarer Rufe (BCI : bad-call-interruption) wird ein Steuerungsaufwand benötigt, der für die Funktionsfähigkeit der BCI-Methode niedrig gehalten werden muß. Der BCI-Steuerungsaufwand wird deshalb in diesem Modellansatz vernachlässigt.

Die hier diskutierte BCI-Methode zur Überlastabwehr basiert auf dem Prinzip, daß die Systemkapazität - in diesem Falle die Teil-

rufverarbeitung - den komplettierbaren Rufen vorrangig zur Verfügung gestellt wird. Zur Erläuterung dieser Methode werden die Zustände und Zustandsübergänge eines Teilrufquellenprozesses in Bild 6.17 mit der SDL-Beschreibungssprache dargestellt (SDL : functional specification and description language [13]).

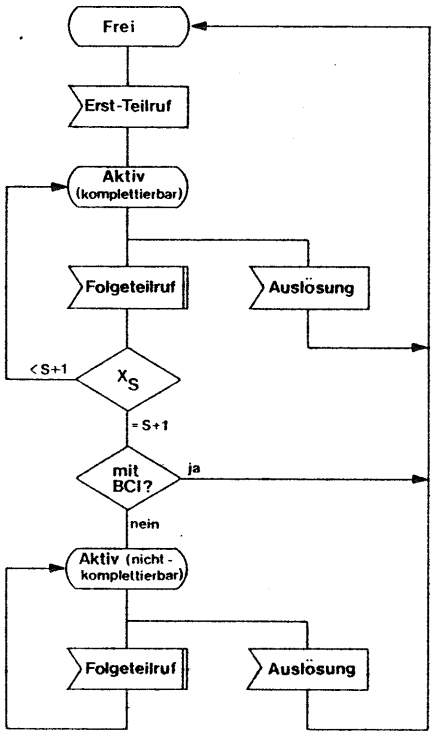


Bild 6.17 *SDL-Zustandsübergangsdiagramm einer Teilrufquelle.*

d) Modellanalyse

Folgende Symbole werden bei der Analyse für die Modellkenngrößen angewendet :

λ	Rufankunftsrate ($\lambda = 1/E[T_{AC}]$)
λ_S	Teilrufankunftsrate einer im aktiven Zustand befindlichen Teilrufquelle ($\lambda_S = 1/E[T_{AS}]$).
ε	Rufenderate ($\varepsilon = 1/E[T_C]$).
μ	Enderate der Teilrufbedienung.
n	maximale Anzahl der Rufe im System.
$\rho_C = \frac{\lambda}{\varepsilon n}$	normiertes Verkehrsangebot der Rufe.
$\rho_S = \frac{\lambda}{\mu} s.n$	normiertes Verkehrsangebot der Teilrufe.
S	Teilruf-Warteschlangenlänge. S+1 ist die Kapazität des Teilruf-Verarbeitungssystems.
Θ	Hilfsparameter zur Kennzeichnung der Überlast-Abwehrstrategie : $\Theta = 0$: ohne BCI. $\Theta = 1$: mit BCI.
X_C	Anzahl der aktiven Teilrufquellen bzw. Rufe.
X_S	Anzahl der Teilrufe im System.
$P(i,j) = P\{X_S=i, X_C=j\}$	Zustandswahrscheinlichkeit für i Teilrufe und j Rufe im System.

Der Systemzustandsprozeß läßt sich durch X_C und X_S vollständig beschreiben. Da alle Ankunfts- und Bedienungsprozesse in diesem Modell die Markoff-Eigenschaft besitzen, kann für die Modellanalyse ein zweidimensionaler Markoff-Prozeß entwickelt werden, dessen Zustandsübergangsdiagramm in Bild 6.18 dargestellt wird.

Anhand einiger exemplarischer Zustände wird im folgenden das Zustandsübergangsdiagramm erläutert. Zunächst werden am Beispiel des Zustandes (X_S=i,X_C=j) die Ereignisse betrachtet, die zu einer

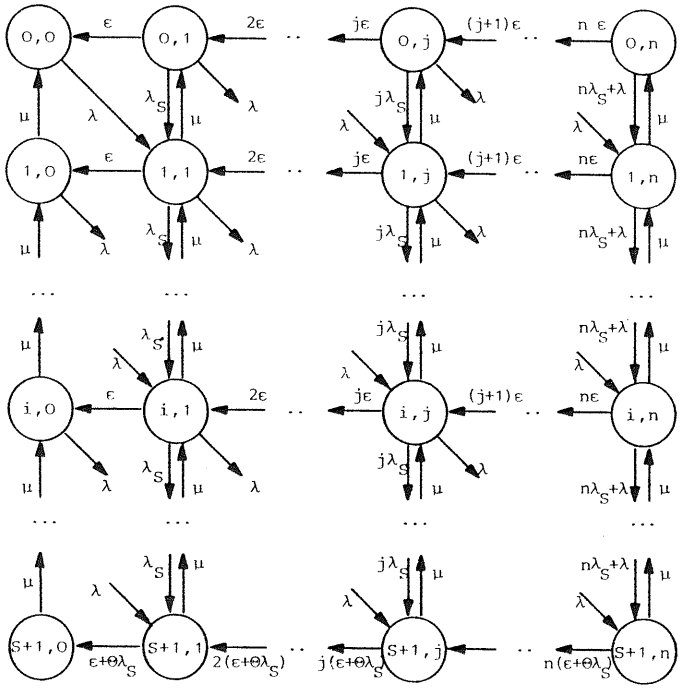


Bild 6.18 *Zustandsübergangsdiagramm des Verkehrsmodells zur Ruf- und Teilrufverarbeitung.*

Zustandsänderung führen können :

- Ankunft eines Rufes mit dem ersten Teilruf. Das System geht mit der Übergangswahrscheinlichkeitsdichte (ÜWD) λ in den Zielzustand (i+1,j+1) über.
- Endigung eines Rufes. Mit der ÜWD $j\varepsilon$ wird der Zielzustand (i,j-1) erreicht.
- Ankunft eines Teilrufs. Der Zielzustand (i+1,j) wird mit der ÜWD $j\lambda_S$ erreicht.
- Bedienungsende eines Teilrufs. Mit der ÜWD μ wird der Zielzustand (i-1,j) erreicht.

Die Zustände (i,n) und (S+1,j), in denen Rufe bzw. Teilrufe blockiert werden, bilden Ausnahmen, die hier kurz diskutiert werden :

- Zustand (i,n) : Rufe werden abgewiesen, da alle Teilrufquellen belegt und aktiv sind. Die ersten Teilrufe werden jedoch angenommen (sofern i<S+1). Der Zustand geht mit der ÜWD λ in den Zustand (i+1,n) über.
- Zustand (S+1,j) : Teilrufe werden abgewiesen, so daß die zugehörigen Rufe nicht mehr komplettiert werden können. Für $\Theta = 1$ geht der Zustand mit der ÜWD $j\lambda_S$ in den Zielzustand (S+1,j-1) über. Für $\Theta = 0$ erfolgt mit der ÜWD $j\lambda_S$ keine Zustandsänderung.

Die Zustandsgleichungen werden gemäß Gl.(4.10) hergeleitet, indem jeder Zustand im statistischen Gleichgewicht betrachtet wird :

$$P(i,j).(\lambda+j\lambda_S+j\varepsilon+\mu) = \lambda P(i-1,j-1) + j\lambda_S P(i-1,j) + (j+1)\varepsilon P(i,j+1) + \mu P(i+1,j) \quad , \quad 0<i<S+1,\ 0\leq j<n, \quad (6.33a)$$

$$P(0,j).(\lambda+j\lambda_S+j\varepsilon) = (j+1)\varepsilon P(0,j+1) + \mu P(1,j) \quad , \quad 0\leq j\leq n, \quad (6.33b)$$

$$P(S+1,j).(j\Theta\lambda_S+j\varepsilon+\mu) = \lambda P(S,j-1) + j\lambda_S P(S,j) + (j+1)(\varepsilon+\Theta\lambda_S)P(S+1,j+1) \quad , \quad 0\leq j<n, \quad (6.33c)$$

$$P(i,n).(\lambda+n\lambda_S+n\varepsilon+\mu) = \lambda P(i-1,n-1) + (n\lambda_S+\lambda)P(i-1,n) + \mu P(i+1,n) \quad , \quad 0<i<S+1, \quad (6.33d)$$

$$P(S+1,n).(n\Theta\lambda_S+n\varepsilon+\mu) = \lambda P(S,n-1) + (n\lambda_S+\lambda)\ P(S,n). \quad (6.33e)$$

Dabei gilt $P(i,-1)=P(-1,j)=P(S+2,j)=P(i,n+1)=0$ für alle i,j.

Zusammen mit der Normierungsbedingung

$$\sum_{i=0}^{S+1} \sum_{j=0}^{n} P(i,j) = 1 \qquad (6.34)$$

bilden die Gleichungen (6.33a-e) ein lineares Gleichungssystem zur Bestimmung der Zustandswahrscheinlichkeiten. Die Berechnung erfolgt mit einem numerischen Iterationsverfahren, da der Zustandsraum in Bild 6.18 keinen einfachen Algorithmus zur rekursiven Auflösung des Gleichungssystems zuläßt.

e) Systemcharakteristiken

Mit den Zustandswahrscheinlichkeiten, die sich aus Gl. (6.33a-e) und Gl. (6.34) numerisch bestimmen lassen, können charakteristische Größen zur Beurteilung der Systemleistung gewonnen werden. Die mittlere Anzahl E_C und E_S von Rufen und Teilrufen im System lautet:

$$E_C = \sum_{i=0}^{S+1} \sum_{j=0}^{n} j\, P(i,j), \qquad (6.35)$$

$$E_S = \sum_{i=0}^{S+1} \sum_{j=0}^{n} i\, P(i,j). \qquad (6.36)$$

Mit Hilfe der Little'schen Formel nach Gl. (4.16) läßt sich die mittlere Durchlaufzeit τ_S eines angenommenen Teilrufes berechnen:

$$\tau_S = \frac{E_S}{\lambda_{AS}}, \qquad (6.37)$$

wobei λ_{AS} die mittlere Rate der angenommenen Teilrufe ist.

$$\lambda_{AS} = \sum_{i=0}^{S} \sum_{j=0}^{n} (j\lambda_S+\lambda)\, P(i,j). \qquad (6.38)$$

Aus Gl. (6.38) ist ersichtlich, daß sich in jedem Zustand (i,j) die Rate λ_{AS} aus zwei Teilraten zusammensetzt: die Rate λ der Erst-Teilrufe und die Rate $j\lambda_S$ der Folgeteilrufe.

Der Quotient der Rate blockierter Teilrufe und aller angebotenen Teilrufe bildet die Blockierungswahrscheinlichkeit B_S für Teilrufe

$$B_S = \frac{\sum_{j=0}^{n} (j\lambda_S+\lambda)\, P\,(S+1,j)}{\lambda_S E_C+\lambda}. \qquad (6.39)$$

Die Blockierungswahrscheinlichkeit für Rufe besteht aus drei Komponenten, die den folgenden Rufblockierungsfällen entsprechen: Rufabweisung wegen besetzter Teilrufquellen, Rufblockierung wegen gefüllter Teilruf-Warteschlange und Zwangsauslösung von Rufen infolge der BCI-Überlastabwehrmaßnahme:

$$B_C = \sum_{i=0}^{S} P(i,n) + \sum_{j=0}^{n} P(S+1,j) + \Theta \frac{\lambda_S}{\lambda} \sum_{j=1}^{n} jP(S+1,j). \qquad (6.40)$$

Speziell für den Fall mit Überlastabwehr (Θ=1) läßt sich die Rufkomplettierungsrate Y bestimmen:

$$Y = \lambda\,(1-B_C). \qquad (6.41)$$

In dem Fall ohne Überlastabwehr kann Y nicht aus den Zustandswahrscheinlichkeiten berechnet werden, da ein Ruf mehrere blockierte Teilrufe erzeugen kann. Die Blockierung eines Folgeteilrufes hat keine Zustandsänderung zur Folge, die zur Berechnung von Y benötigt wird.

Im folgenden wird die Verteilung der Anzahl von Teilrufen pro Ruf hergeleitet, die mit der Zufallsvariable G dargestellt wird. Zuerst wird die bedingte Wahrscheinlichkeit dafür berechnet, daß ein Ruf x Teilrufe produziert, falls die Rufdauer t beträgt:

$$P\{G=x \mid T_C=t\} = \frac{(\lambda_S t)^{x-1}}{(x-1)!} e^{-\lambda_S t}, \quad x = 1,2,\ldots . \tag{6.42}$$

Dabei wird berücksichtigt, daß ein Ruf gleichzeitig mit seinem ersten Teilruf ankommt. Mit Gl. (6.42) und der Verteilungsfunktion für T_C gemäß Gl. (6.31) ergibt sich

$$\begin{aligned} g_x = P\{G=x\} &= \int_0^\infty \frac{(\lambda_S t)^{x-1}}{(x-1)!} e^{-\lambda_S t} \varepsilon e^{-\varepsilon t} dt \\ &= \frac{\lambda_S^{x-1}\varepsilon}{(x-1)!} \int_0^\infty t^{x-1} e^{-(\lambda_S+\varepsilon)t} dt \\ &= \frac{\varepsilon}{\lambda_S} \left(\frac{\lambda_S}{\lambda_S+\varepsilon}\right)^x = \left(1 - \frac{\lambda_S}{\lambda_S+\varepsilon}\right)\cdot\left(\frac{\lambda_S}{\lambda_S+\varepsilon}\right)^{x-1}, \quad x = 1,2,\ldots . \end{aligned} \tag{6.43}$$

Gl. (6.43) stellt eine geometrische Verteilung für die Zufallsvariable G dar. Die mittlere Anzahl der Teilrufe pro Ruf lautet:

$$E[G] = \sum_{x=1}^{\infty} x g_x = 1 + \frac{\lambda_S}{\varepsilon} . \tag{6.44}$$

f) Leistungsbeurteilung der Überlastabwehrmaßnahme

Bild 6.19 zeigt die Teilruf-Blockierungswahrscheinlichkeit als Funktion der Warteschlangenlänge S. In allen Bereichen der Teilruf-Verkehrsintensität λ_S (bzw. ρ_S) läßt sich B_S mit der BCI-Überlastabwehrstrategie reduzieren. Die dazugehörige Rufblockierungswahrscheinlichkeit ist in Bild 6.20 dargestellt. Bei genügend großer Teilruf-Warteschlangenlänge wird die Rufblockierung nur noch von der Anzahl n der Teilrufquellen beeinflußt. Die Rufblockierungswahrscheinlichkeit strebt gegen den Wert des Verlustsystems M/M/n [9].

Die Leistung der BCI-Überlastabwehrstrategie wird in Bild 6.21 illustriert. Die Leistungssenkung des Vermittlungssystems infolge der Bearbeitung nicht-komplettierbarer Rufe kann durch BCI verhindert werden. Im Gegensatz zu den numerisch gewonnenen Ergebnissen mit BCI (Gl.6.41) werden die Kurven ohne BCI (gestrichelte Linien) mit Hilfe eines Simulationsprogramms [48] bestimmt. Die Simulationspunkte sind mit 95%-Vertrauensintervallen aufgetragen.

Parallel zu der Optimierung der Komplettierungsrate reduziert die BCI-Methode die mittlere Durchlaufzeit τ_S der Teilrufe (Bild 6.22). Diese Optimierung rührt daher, daß die BCI-Methode die mittlere effektive Ankunftsrate der Teilrufe durch Zwangsauslösung nicht-komplettierbarer Rufe vermindert.

g) Kombination der BCI-Überlastabwehrstrategie mit vorausschauender Drosselung der Rufannahme

Betrachtet man die in Kap. 3 diskutierten Überlastindikatoren auf der Ruf- und Teilrufebene zusammen, so kann die Leistung der BCI-Überlastabwehrstrategie mit einer vorausschauenden, adaptiven Drosselung der Rufannahme weiter optimiert werden. Dazu werden folgende Überlastindikatoren herangezogen:

- Die Anzahl X_C der aktiven Teilrufquellen bzw. Rufe im System: Mit diesem Indikator kann die Anzahl der in unmittelbarer Zukunft der Prozeßentwicklung zu erwartenden Teilrufe estimiert werden. Durch eine entsprechende Reservierung eines Bereiches der Teilruf-Warteschlange kann die Teilrufblockierung reduziert und dadurch die Komplettierungswahrscheinlichkeit der bereits angenommenen Rufe erhöht werden.
- Die Anzahl X_S der Teilrufe im System: Dieser Indikator zeigt den aktuellen Belastungszustand der Teilrufverarbeitung an.

Folgende Rufannahmestrategie kann zusammen mit der BCI-Strategie kombiniert werden. Danach werden angebotene Rufe vorsorglich abgewiesen, wenn:

$$K_C X_C + K_S X_S > S_O . \tag{6.45}$$

Dabei sind K_C der rufbezogene und K_S der teilrufbezogene Gewichtungsfaktor der Überlastindikatoren. Der Schwellenwert S_O ($O \leq S_O \leq S$) markiert die Grenze eines reservierten Bereiches ($S-S_O$) in der Warteschlange für die Teilrufe, die von komplettierbaren Rufen erzeugt werden.

Die hier diskutierte Kombination der BCI-Überlastabwehrstrategie mit einer vorausschauenden Drosselung der Rufannahme wird in [48] simulativ untersucht. Dieser Simulationsstudie folgen Untersuchungen analytischer Art zur Leistungsbeurteilung der Überlastabwehrstrategie, die im Rahmen der vorliegenden Arbeit nicht weiter behandelt werden.

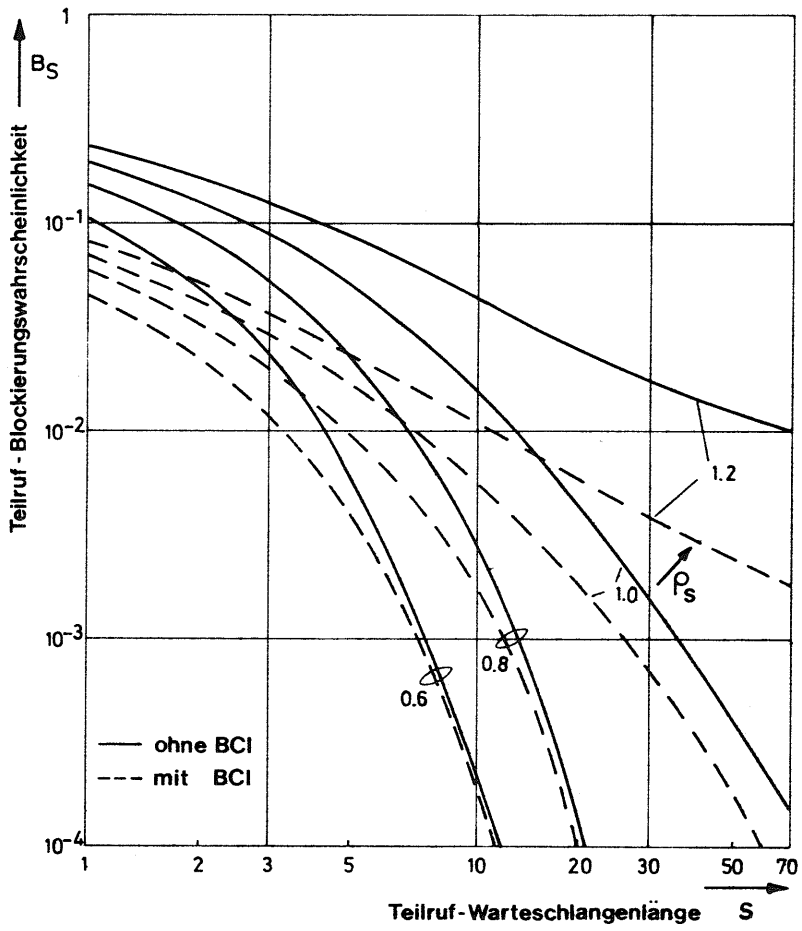


Bild 6.19 *Reduzierung der Teilruf-Blockierungswahrscheinlichkeit mittels BCI-Überlastabwehrstrategie.*

Parameter : $\rho_C = 0.7$, $n = 50$, $E[G] = 21$

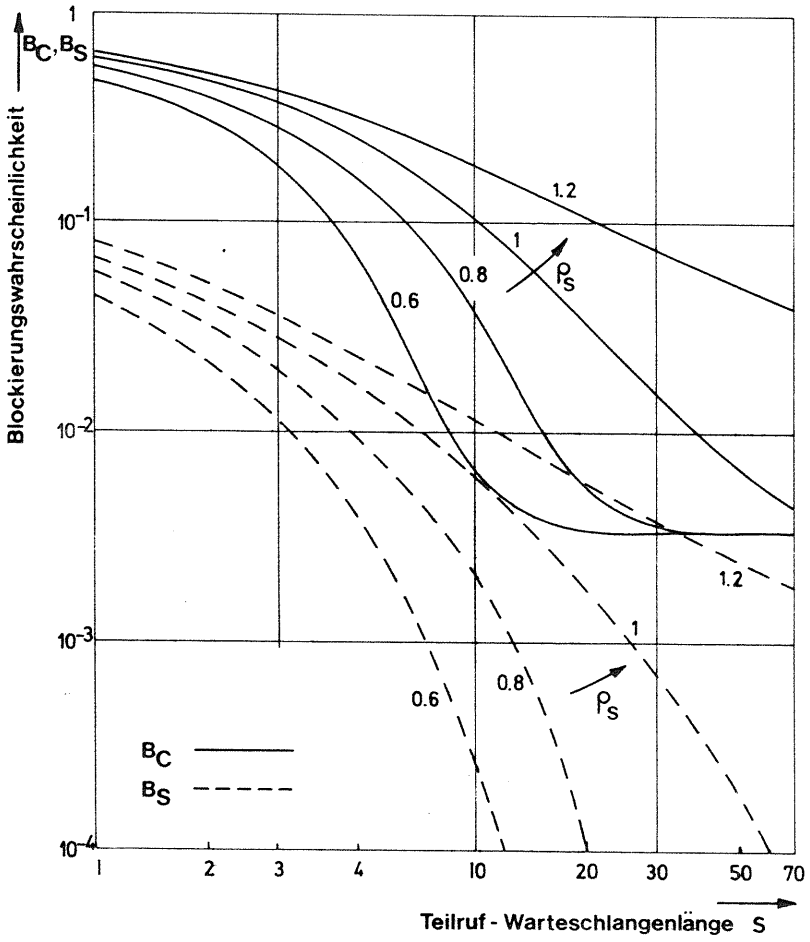


Bild 6.20 *Ruf- und Teilrufblockierung mit der BCI-Überlastabwehrstrategie.*

Parameter : $\rho_C = 0.7$, $n = 50$, $E[G] = 21$

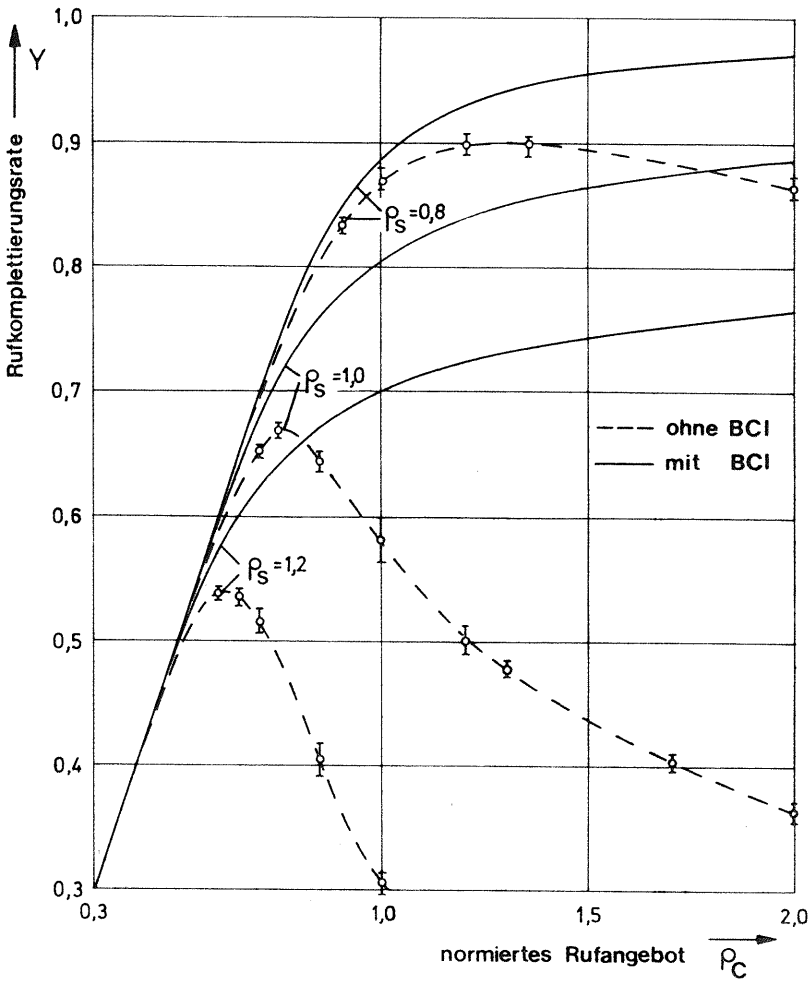


Bild 6.21 *Optimierung der Rufkomplettierungsrate mit der BCI-Überlastabwehrstrategie.*

Parameter : $S = 20$, $n = 50$, $E[G] = 21$

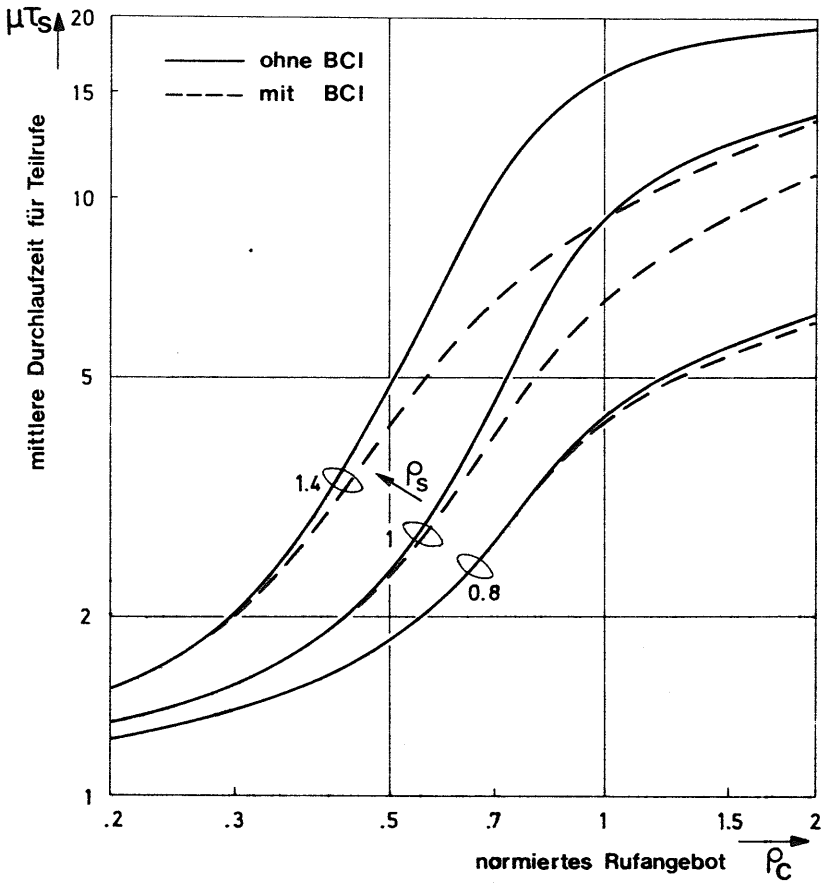


Bild 6.22 *Reduzierung der mittleren Teilruf-Durchlaufzeit mittels BCI- Überlastabwehrstrategie.*

Parameter : S = 20 , n = 50 , E[G] = 21

7. ZUSAMMENFASSUNG

In der vorliegenden Arbeit wurden Überlast und Überlastabwehrproblematik in rechnergesteuerten Fernsprechvermittlungssystemen behandelt. Aspekte der statistischen Beschreibung und Erfassung von Überlastsituationen sowie der Leistungsbeurteilung von Überlastabwehrmaßnahmen wurden mit Methoden zur verkehrstheoretischen Modellbildung und -analyse untersucht.

Nach einem Überblick über Steuerungsstrukturen rechnergesteuerter Fernsprechvermittlungssysteme wurden Ursachen und Indikatoren der Systemüberlastung sowie darauf aufbauende Überlastabwehrstrategien in einer klassifizierenden Form diskutiert. Dabei sind strukturelle und steuerungstechnische Systemmerkmale sowie das Teilnehmerverhalten einbezogen worden.

Analytische und simulative Methoden zur stationären und instationären Modellanalyse wurden diskutiert, wobei u.a. eine neue Methode zur Simulation instationärer Zufallsprozesse, insbesondere des in Überlastuntersuchungen häufig verwendeten verallgemeinerten Poisson-Prozesses, vorgestellt wurde.

In den abschließenden Kapiteln wurden Verkehrsmodelle entwickelt und analysiert, mit denen

- Verkehrsströme in Überlastsituationen bzw. die dynamische Entwicklung von Überlast beschrieben werden kann.
- eine quantitative und qualitative Leistungsbeurteilung entwickelter Überlastabwehrstrategien ermöglicht wird. Hierbei wurden u.a. mit analytischen Mitteln die stochastischen Systemantworten auf kurzzeitige Überlastimpulse bestimmt, welche bei instationären Überlastabwehrvorgängen typisch sind.

Aufgrund der immer komplexer werdenden Systemstruktur und -organisation rechnergesteuerter Fernsprechvermittlungssysteme ist es notwendig, einen leistungsfähigen, systemspezifischen Überlaststeuerungsplan für das ganze System zu entwickeln. Anhand einer

differenzierten Verwendung mehrerer Überlastindikatoren kann eine gezielte Aktivierung der im Überlaststeuerungsplan vorgesehenen, hinsichtlich der Wirkgeschwindigkeit und der Wirkbreite abgestuften Überlastabwehrmaßnahmen erfolgen.

Die in der vorliegenden Arbeit vorgestellten Verkehrsmodelle und Modelluntersuchungen ermöglichen die qualitative und quantitative Beurteilung derartiger Regelungsmechanismen.