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

38. Bericht über verkehrstheoretische Arbeiten

Modellierung und Analyse von Überlast-Abwehrmechanismen in Paketvermittlungsnetzen

von
Harmen R. van As

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

38th Report on Studies in Congestion Theory

Modelling and Analysis of Congestion Control Mechanisms in Packet Switching Networks

by
Harmen R. van As

1984

Druck: E. Kurz & Co., Stuttgart

ISBN 3-922403-48-4

ABSTRACT

This research report deals with the performance evaluation of congestion control mechanisms in packet switching networks. First, an introduction on typical architectures of packet switching networks and their operational principles will be given. Then, causes of and indicators for congestion are discussed, and a classification of congestion control mechanisms is presented.

Queueing models have been used for the quantitative consideration of congestion and for the performance evaluation of adequate congestion control mechanisms. To deal with the dynamic character of congestion, several queueing models have been analyzed for the nonstationary case. Adequate mathematical tools and their computational implementation will be described. In addition, a flexible and powerful simulation system is introduced.

In a first category of queueing models, typical forms of congestion are analyzed: the propagation of congestion, backpressure, network breakdown, and priority deadlock.

The second category of queueing models describes congestion control mechanisms: network access, the foreground-background strategy, two-point control, and adaptive time-outs.

CHAPTER 1 : Introduction - Overview (pp. 5-7)

The first chapter gives the motivations for the wide-spread introduction of packet switching and indicates the inherent danger of sudden congestion. In addition, an overview of the content of this report is given.

CHAPTER 2 : Packet switching networks (pp. 8-29)

In this chapter, the principles of packet switching, the typical structures for networks and nodes, and the different modes of operation are described.

Furthermore, considerable space has been devoted to different aspects of the related protocols: protocol hierarchy, standards, flow control, automatic error correction, and routing.

Section 2.1 deals with the basic characteristics and features of packet switching. Section 2.2 describes the structure of current packet switching networks and their connection to other communication networks. Section 2.3 characterizes the architecture and functions of modular multi-processor switching nodes. Section 2.4 compares the basic modes of operation: datagram and virtual circuit. Section 2.5 emphasizes the central role of communication protocols and the current efforts for open system interconnection. Section 2.6 gives a summary of the ISO-protocol architecture. Section 2.7 deals with relevant CCITT recommendations and their relationship. section 2.8 lists the features of the recommendation X.25 and illustrates the different phases of a data exchange. Section 2.9 describes the functions of flow control at the different protocol levels. Special attention has been paid to the concept of window flow control. Section 2.10 is devoted to automatic error correction. The causes triggering a recovery are listed and the procedures described briefly. Section 2.11 gives an overview on routing algorithms.

CHAPTER 3 : The congestion problem (pp. 30-50)

Congestion in packet-switching networks is strongly related to the interaction between stochastic demand and the limited number of service facilities.

Therefore, a complete investigation on congestion control comprises the following: characterization of the traffic offered, traffic-related aspects for implementations, causes of and indicators for congestion, and classification and description of congestion control mechanisms.

Section 3.1 compares the traffic characteristics of interactive and batch applications. Section 3.2 discusses the impact of implementation aspects: acknowledgement delay, throughput, fairness, utilization, availability, network access, reliability, and economics. Section 3.3 characterizes a congested network (throughput, delay), deals with primary and secondary causes for this situation, and discusses indicators for early detection. Section 3.4 describes some implementation aspects

of congestion control mechanisms: hierarchical architecture, radius and speed of operation, and the way of realization. For this last aspect, a detailed classification has been made: storage management, flow control, scheduling, routing, topology, and tariff policy. Section 3.5 refers to similar congestion problems in telephone networks.

CHAPTER 4 : Modelling and analysis (pp. 51-103)
In this chapter, traffic-theoretic modelling and its analysis are given. In particular, the theory of transient queueing analysis is described, extended, and implemented.
With respect to simulation, a flexible and powerful tool is introduced.

Section 4.1 gives the elements for traffic-theoretic modelling of packet switching networks: model structure, mode of operation, traffic characteristics and parameters, and evaluation criteria. Section 4.2 deals with the description of queueing models by time-dependent Markovian processes: system state process, waiting process, and flow process. Furthermore as an example, the time-dependent characteristic traffic values are given for the basic queueing system M/M/1/S. Section 4.3 surveys the theory of queueing networks with product form solution and its applicability to packet switching networks. Section 4.4 describes numerical methods to solve Markovian queueing models: stationary (SOR-point-iteration, recursive solution technique) or nonstationary (4th-order Runge-Kutta method). Section 4.5 summarizes the principles of stochastic traffic simulations both for the stationary and the nonstationary case. Section 4.6 introduces two program systems: a program system for the general numerical solution of complex Markovian queueing models, and the simulation package QSIMLIB supporting development of stationary or nonstationary simulation programs for complex queueing systems or networks.

CHAPTER 5 : Modelling of congestion (pp. 104-135)
This chapter deals with queueing models describing typical situations of congestion. These models contribute to a thorough understanding of the nature of congestion.

Section 5.1 considers N Markovian queueing systems in series, and demonstrates the propagation speed and the aftereffect of a rectangular overload peak. Section 5.2 makes use of a coupled system of two finite Markovian queueing systems to demonstrate the back-pressure effect in case of a congested second system. Section 5.3 illustrates, by simulation, the sudden system breakdown when an uncontrolled network is flooded with packet copies generated automatically upon acknowledgement delay time-out. Section 5.4 demonstrates the time-dependent behaviour of a finite Markovian queueing system with two traffic streams when one of the streams generates an overload peak. The case of nonpreemptive priorities can lead to a so-called priority deadlock.

CHAPTER 6 : Network access control (pp. 136-153)
An effective way to prevent a congested network is to control network access. This chapter describes and analyzes this class of congestion control mechanisms.

Section 6.1 describes the background and existing literature of this class of congestion control mechanisms. Section 6.2 defines the corresponding Markovian queueing model with two traffic streams with different conditions for acceptance. The case of nonpreemptive priorities will be considered as well. Section 6.3 gives the description of the analysis, and defines the characteristic traffic values for both traffic classes: loss probability, mean system occupancy, and throughput. Section 6.4 shows the corresponding stationary and nonstationary results. Section 6.5 contains the conclusion that both static and dynamic results show that the network-access control mechanisms considered here, which can easily be implemented, work efficiently.

CHAPTER 7 : Foreground-background control (pp. 154-193)
Whereas packets entering the network can be rejected without waste of network resources, a rejection of packets already within the network should be avoided. For these so-called transit packets a congestion control mechanism based on priority scheduling and dynamic foreground-background storage management is proposed and analyzed.

Section 7.1 describes the motivation for the foreground-background control strategy: each network node should try to manage a short overload peak by itself instead of shifting the congestion to the neighboring nodes. Section 7.2 defines the underlying Markovian queueing model consisting of a finite foreground store, a large background store, and two nonpreemptive priority traffic streams with different acceptance rules. Section 7.3 deals with the analysis on the basis of the corresponding system state process. Section 7.4 considers the flow process of the model. Section 7.5 summarizes all relevant characteristic traffic values for both priority classes: loss probability, mean system occupancy, throughput, and mean flow time. Section 7.6 shows stationary and nonstationary results. Section 7.7 states that in a packet switching network with two priority classes (e.g., dialog and batch) the FG-BG strategy is able to cope effectively with short overload peaks, and that it can be considered as an important part of an overall congestion control plan.

CHAPTER 8 : Two-point control (pp. 194-221)
In the previous two chapters, congestion control mechanisms which work locally have been analyzed. This chapter deals with a global mechanism: Messages to the data sources initiate adequate slowdown activities.

Section 8.1 motivates the need for this kind of congestion control and describes the two-point control mechanism. Section 8.2 defines the Markovian queueing models: interrupted control and alternating control, both with delayed reaction. Section 8.3 describes the analysis of the queueing models. For

the interrupted control model, the analysis is based either on a simpler or on a complex state description. Furthermore, the characteristic traffic values have been defined in this section: loss probabilities, mean system occupancy, probability for lost capacity. Section 8.4 shows stationary and nonstationary results. Section 8.5 emphasizes that the two-point control mechanisms works well as long as the reaction delay remains short. Moreover, some optimization criteria are given.

CHAPTER 9 : Adaptive time-out control (pp. 222-245)
This chapter addresses design and operational guidelines for packet retransmission strategies necessary to support recovery from lost or damaged packets.

Section 9.1 gives an introduction to the problem of selecting an optimal time-out interval. Section 9.2 discusses aspects related to time-out management: length of the time-out interval, number of packet copies allowed, procedure in the case of a time-out event. In addition, the concept of the time-out window is presented. Section 9.3 deals with the modelling part. First, the basic model for the time-out mechanism is defined. Thereafter, several models are described: an analytical model based on the determination of a distribution function, an analytical model based on a process description, and simulation models. Section 9.4 gives numerical results: comparison between the different methods, basic strategy for adaptive time-outs, strategy modifications, and time-out window control. Section 9.5 concludes that the problem of unnecessary packet copies can be managed completely by a combination of adaptive time-outs and time-out window control.

CHAPTER 10 : Summary (pp. 246-247)
In this chapter, a summary is given. Moreover, the importance of considering dynamic aspects in the analysis of congestion control mechanisms is pointed out.

INHALTSVERZEICHNIS

1. EINLEITUNG - ÜBERSICHT ÜBER DIE ARBEIT

Die Fortschritte in der Mikroelektronik, insbesondere auf dem Gebiet der Mikroprozessoren und Speicher, ermöglichen heute eine kostengünstige Realisierung für die verteilte Verarbeitung von Daten. Aus der Vielfalt der Anwendungsmöglichkeiten sind zum Beispiel zu nennen: Reservierungssysteme, Abwicklung von Bankverkehr, Zugriff auf Datenbanksysteme sowie Kommunikation zwischen Rechenzentren. Grundlegend für ihre Verwirklichung ist ein leistungsfähiges und zuverlässiges Datennetz, entweder als Leitungsvermittlungsnetz oder als Paketvermittlungsnetz. Während bei der Leitungsvermittlung den beteiligten Endeinrichtungen für die Dauer der Verbindung ein durchgehender physikalischer Übertragungsweg für die Nachrichten zur Verfügung gestellt wird, teilen sich die sogenannten logischen Verbindungen bei Paketvermittlung die Übertragungswege im asynchronen Zeitmultiplex. Ähnlich wie in einem Straßennetz entstehen durch statistische Verkehrsschwankungen unerwartete Stausituationen, die eine Verringerung des Datendurchsatzes und eine Erhöhung der Netzdurchlaufzeiten zur Folge haben. Bedingt durch die Paketwiederholmechanismen kann das Netz sogar zusammenbrechen. Überlastabwehr in Fernsprech- und Paketvermittlungsnetzen gilt seit einigen Jahren als aktuelles Forschungsgebiet, und wichtige Ergebnisse sind bereits erzielt worden. Die Komplexität dieses Themas läßt aber noch eine Fülle von theoretischen und praktischen Fragen unbeantwortet. Diese Arbeit soll dazu beitragen, weitere Erkenntnisse in der Modellierung und Untersuchung von Überlastabwehrstrategien in Paketvermittlungsnetzen zu gewinnen.

Im Kapitel 2 wird ein kurzer Überblick über typische Strukturen, Betriebsarten und Ablaufvorgänge in Paketvermittlungsnetzen gegeben. Besondere Aufmerksamkeit wird den Protokollen gewidmet. Dies betrifft ihre hierarchische Struktur gemäß dem ISO-Architekturmodell, die Datenflußsteuerung sowie die automatischen Fehlerkorrekturverfahren. Für Paketvermittlungsnetze relevante Empfehlungen nach CCITT werden ebenfalls kurz beschrieben.

Kapitel 3 behandelt Ursachen und Indikatoren für Überlastsituationen in Paketvermittlungsnetzen. Ferner werden die verschiedenen Abwehrmaßnahmen zusammengestellt und diskutiert.

In Kapitel 4 werden die verkehrstheoretische Modellbildung beschrieben und die in dieser Arbeit generell angewandten Analyseverfahren betrachtet. Spezielle Analysemethoden werden jeweils an der betreffenden Stelle behandelt. In diesem Kapitel wird ferner ein Programmsystem zur Lösung von mehrdimensionalen Markoff-Prozessen unter stationären oder transienten Bedingungen vorgestellt. Einige Modelle werden mit Hilfe von stochastischen Simulationen untersucht. Die diesbezügliche Programmerstellung wird von einem leistungsfähigen und flexiblen Simulationssystem unterstützt.

In den weiteren Kapiteln wird eine Reihe von Überlastabwehrstrategien untersucht, die je nach Netzimplementierung bereits bestehende Mechanismen erweitern sollen. Hierbei gilt allgemein, daß Stausituationen in Paketvermittlungsnetzen nur durch eine sinnvolle Kombination und Stufung von verschiedenen Überlastabwehrmaßnahmen zu bewältigen sind. Voraussetzung für ein wirksames Zusammenspiel ist das Verständnis jedes einzelnen Mechanismus. Die nachfolgenden Kapitel sollen dazu beitragen.

Kapitel 5 befaßt sich mit Modellen, die das Zustandekommen von typischen Überlastsituationen in Paketvermittlungsnetzen charakterisieren sollen. Es handelt sich hierbei insbesondere um die Ausbreitung von Überlastsituationen, den Rückstau von Paketen, die lawinenartige Überflutung des Netzes mit Paketkopien und den Einfluß der begrenzten Speicherkapazität.

In Kapitel 6 werden verschiedene Strategien zur Regelung des Netzzugangsverkehrs miteinander verglichen. Als Entscheidungskriterium zur Abweisung von Paketen aus dem Anschlußnetz ist der momentane Belegungszustand im betrachteten Netzknoten maßgebend.

In Kapitel 7 wird eine Überlastabwehrstrategie untersucht, in der die Netzknoten in der Lage sind, Pakete mit unkritischen Durchlaufzeitanforderungen, wie dies bei Stapelbetrieb der Fall ist, für eine kurze Zeit auszulagern. Auf diese Weise kann Speicherplatz für Pakete mit Realzeitbedingungen (Dialog, Netzsteuerung) je nach momentanem Bedarf freigemacht werden. Kapazitätsmindernde Paketwiederholungen können dadurch vermieden werden, und so wird der Stauausbreitung auf andere Netzknoten entgegengewirkt.

Kapitel 8 behandelt eine Begrenzung der exzessiven Ankunftsrate von Paketen mittels einer Zweipunkt-Regelung. Diese Regelung ist entweder intermittierend, oder es wird zwischen zwei verschiedenen Ankunftsraten hin- und hergeschaltet. Dabei wird die stochastische Verzögerung bis zum Einsetzen der Regelung berücksichtigt.

Kapitel 9 befaßt sich mit dem Problem der Überflutung des Netzes mit Paketkopien infolge von Quittierungsverzögerungen. Als Gegenmaßnahme wird die Zeitüberwachung adaptiv eingestellt. In einer erweiterten Strategie wird die Paketwiederholung bei einer momentan hohen Netzbelastung verzögert, um eine weitere Erhöhung der Belastung zu verhindern.

2. AUFBAU UND BETRIEB VON PAKETVERMITTLUNGSNETZEN

2.1 Prinzip der Paketvermittlung

Im Bereich der Datenübermittlung hat das Prinzip der Paketvermittlung in den letzten Jahren zunehmend an Bedeutung gewonnen. Bei diesem Übermittlungsverfahren werden die in digitaler Form vorliegenden Nachrichten zunächst in Pakete mit variabler, aber beschränkter Länge unterteilt. Jedes Paket enthält neben der Teilnachricht noch einen sogenannten Paketkopf mit Adress- und Steuerinformationen. Die Pakete werden dann über den nächstliegenden Netzknoten dem Paketvermittlungsnetz übergeben. In jedem Netzknoten werden die eintreffenden Pakete zunächst zwischengespeichert, anhand der Information im Paketkopf identifiziert und dann zum nächsten Netzknoten weitergeleitet, sobald der betreffende Übertragungskanal frei ist.

Die wesentlichen Vorteile der Datenpaketvermittlung gegenüber herkömmlichen Systemen, die nach dem Prinzip der Leitungsdurchschaltung arbeiten, liegen zum einen in der besseren Kanalausnutzung und zum anderen in der größeren Flexibilität. Dabei ist die bessere Kanalausnutzung auf die Tatsache zurückzuführen, daß nur dann Kanalkapazität beansprucht wird, wenn Daten zu übertragen sind. In den Pausen steht die Kapazität anderen Benutzern zur Verfügung. Dieses Verfahren ist deshalb für alle Anwendungen günstig, bei denen Datenübermittlungen mit häufig wechselnder Intensität vorkommen, wie dies zum Beispiel bei einer interaktiven Rechnerbenutzung der Fall ist. Die größere Flexibilität beruht auf der relativ einfach zu realisierenden Möglichkeit, in den Netzknoten Geschwindigkeits-,Protokoll- oder Codeumwandlungen vornehmen zu können, so daß auch zwischen zueinander inkompatiblen Endgeräten ein Datenaustausch möglich wird. Die blockweise Übermittlung von Nachrichten erlaubt ferner eine abschnittsweise Fehlersicherung mittels zyklischer Codes zur Fehlererkennung und anschließender Fehlerkorrektur.

Eine historischer Abriß ist in [Roberts (1978)] zu finden. Neuzeitliche Bücher über die Paketvermittlung sind [Schnupp (1978), Davies/Barber/Price/Solomonides (1979), Kerner/Bruckner (1981), Tanenbaum (1981)].

2.2 Struktur von Paketvermittlungsnetzen

In einem Paketvermittlungsnetz (Bild 2.1) unterscheidet man zwischen Anschlußnetz und Vermittlungsnetz. Das Vermittlungsnetz besteht aus den Netzknoten zur Vermittlung der Pakete, den Verbindungsleitungen und einem Netzkontrollzentrum, das zentrale Steuerungs- und Überwachungsfunktionen ausübt und die Gebühreninformationen sammelt. Abhängig von geographischer Struktur, Netzverfügbarkeitsbedingungen und Verkehrsvolumen wird eine kostengünstige, vermaschte Netztopologie ausgewählt [Kleinrock (1976), Schwartz (1977), Garcia (1982)]. Größere Netze werden hierarchisch aufgebaut.

Zu dem Anschlußnetz gehören die Datenendeinrichtungen und deren Anschlußleitungen. Die paketorientierten Datenstationen und Datenverarbeitungsanlagen können über eine genormte Schnittstelle angeschlossen werden. Für zeichenorientierte Datenstationen oder Anschlüsse über das Fernsprechwählnetz und das leitungsvermittelnde Datennetz sind als PAD-Funktion (Packet Assembly/Disassembly) bezeichnete zusätzliche Einrichtungen notwendig. Sie nehmen die Umwandlung in eine paketierte Darstellung und umgekehrt vor. Verbindungen mit anderen Paketvermittlungsnetzen sind über sogenannte Gateways möglich. Angaben über das Paketvermittlungsnetz DATEX-P der Deutschen Bundespost findet man in [Gabler/Tietz (1981), Hillebrand (1981)].

2.3 Aufbau und Funktion der Netzknoten

Moderne Netzknoten sind charakterisiert durch eine modulare Multiprozessor-Architektur: alle Systemeinheiten sind völlig autonom und über ein aus Sicherheitsgründen gedoppeltes Bussystem miteinander verbunden [Bux/Kühn/Kümmerle (1979), Druckarch/van den Burg (1980), Sproule/Mellor (1981)]. Bild 2.2 zeigt eine typische Systemstruktur bestehend aus:

- Vermittlungseinheiten für die Verarbeitung der Adress- und Steuerinformationen der einzelnen Pakete,
- Leitungseinheiten für die fehlerfreie Paketübertragung über die Teilnehmeranschlußleitungen sowie über die Netzverbindungsleitungen,

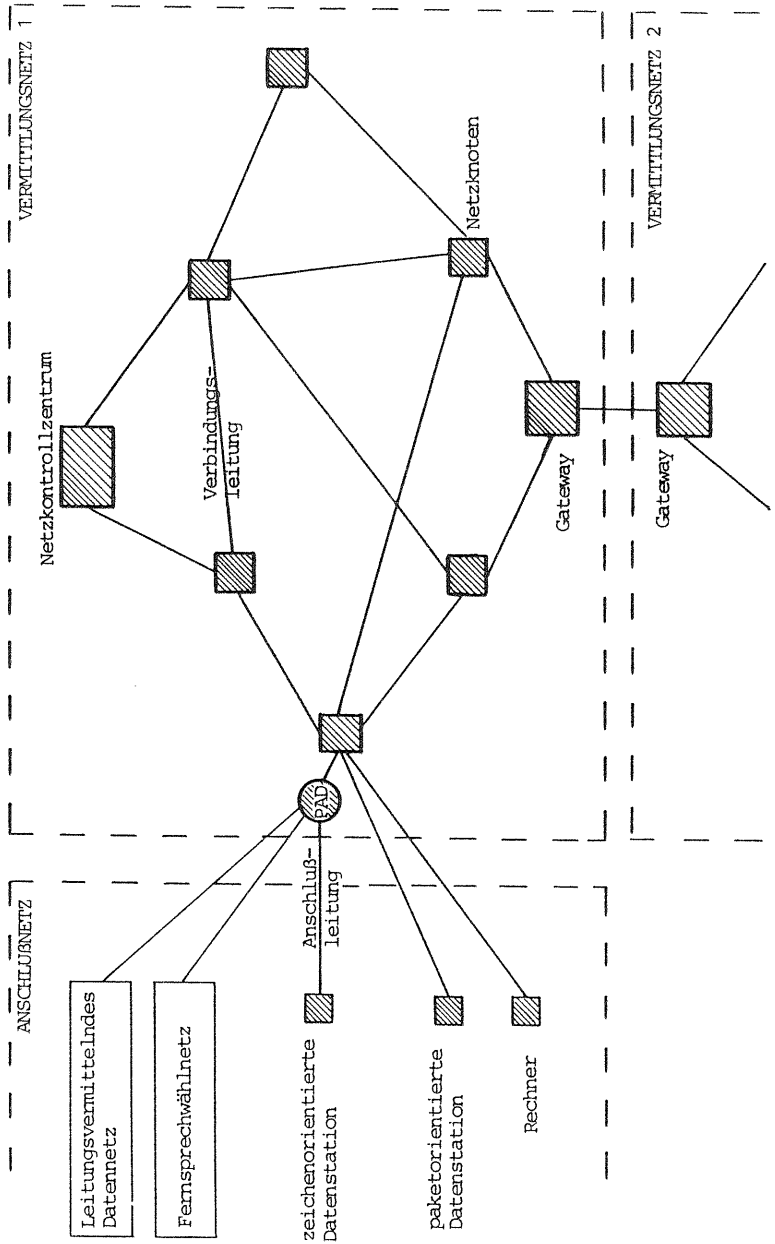


Bild 2.1: Struktur eines Paketvermittlungsnetzes

- eine Knotenüberwachungseinheit für die übergeordneten Überwachungs- und Steuerungsfunktionen,
- und eine Knotendatenbank zur Verwaltung der Teilnehmerdaten (Adressen, Berechtigung, Gebühren).

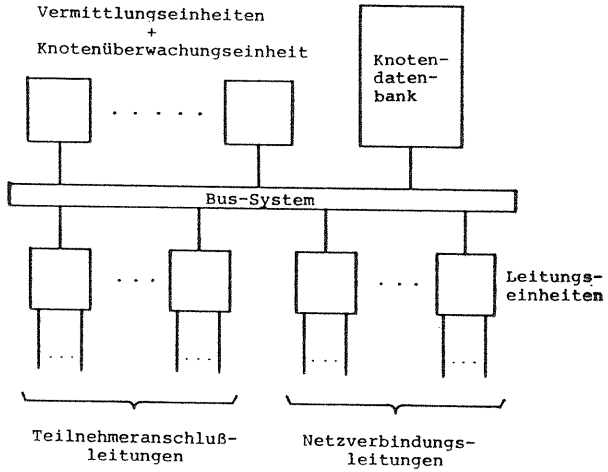


Bild 2.2: Struktur eines Netzknotens mit autonomen Systemeinheiten.

2.4 Betriebsarten

Im wesentlichen haben sich zwei Betriebsarten und ihre Kombination durchgesetzt: die virtuelle Verbindung und das Datagramm [Folts (1980), Rybczynski (1980)].

Unter einer virtuellen Verbindung versteht man eine temporäre logische Beziehung zwischen den beteiligten Datenendeinrichtungen. Insbesondere wird garantiert, daß die gesendeten Pakete fehlergesichert und in der Sendereihenfolge zum Zielteilnehmer gelangen. Auf diese Weise werden den Datenteilnehmern, abgesehen von einer variablen Verzögerung, die Eigenschaften einer real durchgeschalteten Verbindung mit Fehlersicherung geboten. Die virtuelle Verbindung muß vor Beginn der Daten-

Übertragung mit Hilfe speziell gekennzeichneter Pakete aufgebaut werden und wird nach der Übertragung wieder abgebaut. Für Teilnehmer, die sehr häufig miteinander in Verbindung stehen, ist eine permanente logische Verbindung vorteilhaft. Man spricht in diesem Falle von einer festen virtuellen Verbindung. Bei Transaktionen ist der Betriebsmodus "Einzelpaket" (fast select) sinnvoll: Pakete zum Aufbau der virtuellen Verbindung können gleichzeitig Informationen mitführen und mit dem entsprechenden Antwortpaket wird die Verbindung wieder abgebaut.

In einem reinen Datagramm-Netz werden die von einer Datenendeinrichtung in das Netz abgegebenen Pakete aufgrund der in jedem Paketkopf enthaltenen Adresse unabhängig voneinander durch das Netz zum Ziel transportiert. Das Netz garantiert aber nicht die korrekte Übermittlung aller abgegebenen Datagramme, und im besonderen gewährleistet es nicht die Beibehaltung ihrer zeitlichen Reihenfolge. Die Überprüfung der korrekten Zustellung und eventuell die Wiederherstellung der Sendereihenfolge ist eine Aufgabe, die Sender und Empfänger selbst übernehmen müssen.

Als Vorteile der Datagramm-Betriebsart gegenüber derjenigen mit virtuellen Verbindungen gelten eine einfachere Netzanschlußschnittstelle, einfachere Prozeduren innerhalb des Netzes, kürzere Durchlaufzeiten für Einzelpakete und ein einfacherer Übergang zu anderen Netzen. In einigen Paketvermittlungsnetzen hat man die Vorteile beider Verfahren kombiniert, indem zwischen dem Netzknoten und der Datenendeinrichtung eine virtuelle Verbindung realisiert wird, während die Pakete innerhalb des Vermittlungsnetzes nach dem Datagramm-Verfahren transportiert werden. Dies ist der Fall im kanadischen Netz DATAPAC bzw. im DATEX-P [Sproule/Mellor (1981)] und im niederländischen Netz DN1 [Soto/Miguez/Niemegeers (1983)].

2.5 Protokolle

Zur Gewährleistung eines zuverlässigen Informationsaustausches zwischen Datenendeinrichtungen ist eine Vielzahl von Funktionen notwendig. Die Regeln, denen diese Funktionen gehorchen, werden als Protokolle bezeichnet. Sie bestimmen den zeitlichen Ablauf

und die Paketformate für den Informationsaustausch. Die Komplexität dieser Protokolle und die Forderung an Flexibilität, Erweiterbarkeit und nicht zuletzt Übersichtlichkeit bedingen eine modulare Aufteilungsstruktur in hierarchisch gegliederte Funktionsbereiche. Somit entstehen sogenannte Protokollebenen. Währenddem in privaten Netzen oder in Netzen eines einzigen Herstellers eine individuelle Funktionseinteilung möglich ist, sind für Datennetze, die eine herstellerunabhängige Kommunikation (Open Systems Interconnection) unterstützen sollen, die einzelnen Funktionen jeder Protokollebene und deren Schnittstellen zu normieren.

2.6 Das ISO-Architekturmodell

Bild 2.3 zeigt das von der International Standards Organisation (ISO) definierte Architekturmodell, das in den weltweiten Normierungsaktivitäten als Referenz dienen soll [DP-ISO-Basic Reference Model]. Dieses Modell besteht insgesamt aus 7 Ebenen, wobei unterschieden wird zwischen den Ebenen der Transportfunktionen (Ebenen 1-4) und denen der Benutzerfunktionen (Ebenen 5-7).

Je nach Ebene spricht man einerseits von Bit (Ebene 1), Rahmen (Ebene 2), Paket (Ebene 3), oder Nachricht (Ebenen 4-7), andererseits allgemein von Block unter Angabe der betreffenden Ebene.

Im Einzelnen sind die notwendigen Aufgaben wie folgt über die Protokollebenen verteilt:

- Ebene 1: Die Bitübertragungsebene stellt die mechanischen, elektrischen, funktionellen und prozeduralen Eigenschaften bereit, die für Aufbau, Überwachung und Auslösung von physikalischen Verbindungen zwischen Datenendeinrichtung und Netzknoten oder zwischen Netzknoten untereinander sowie die Übertragung eines bitseriellen Datenstromes benötigt werden.
- Ebene 2: Die Sicherungsebene ermöglicht eine zuverlässige Übertragung über einen einzigen Übertragungsabschnitt und umfaßt die Rahmenverwaltung, die Flußsteuerung auf dem Abschnitt sowie die Prozeduren des Aufbaus und der Auslösung von Verbindungen über den Abschnitt.

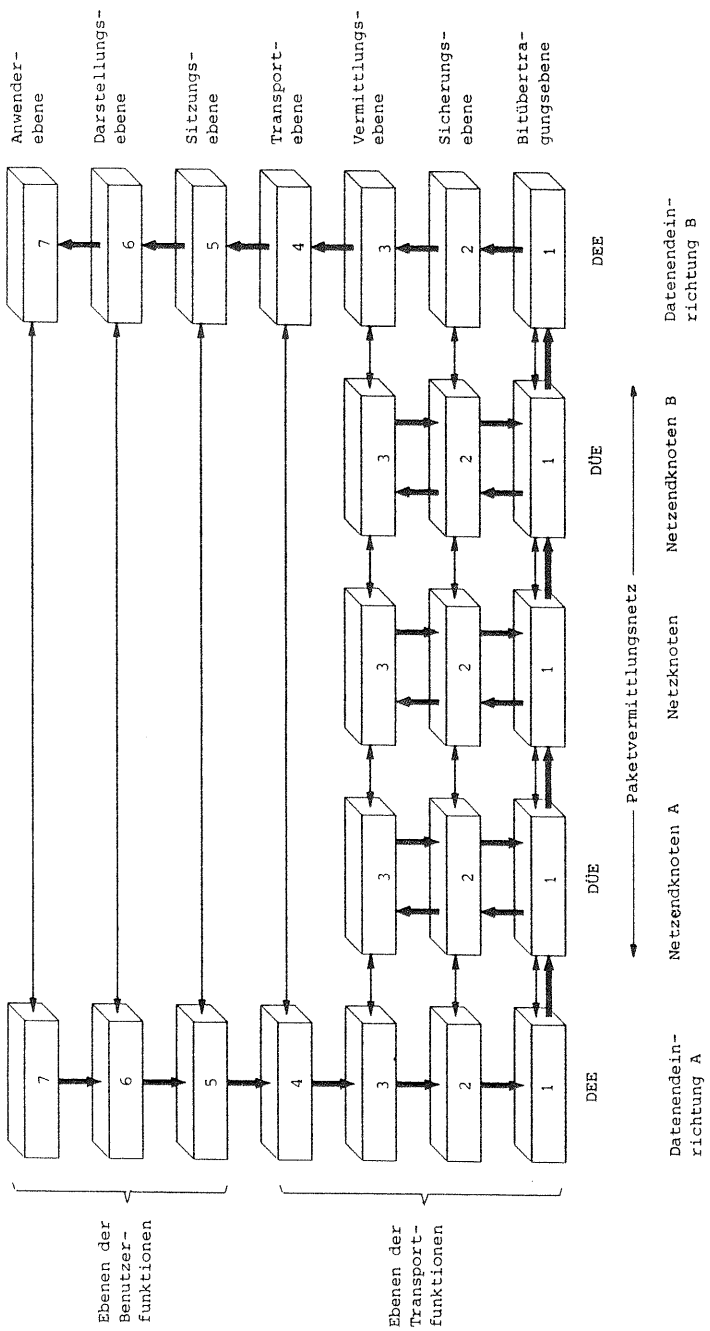


Bild 2.3: ISO-Architekturmodell.

- Ebene 3: Die Vermittlungsebene besorgt den netzweiten Pakettransport vom Ursprungsanschluß zum Zielanschluß. Um dies zu ermöglichen, übernimmt diese Ebene Funktionen wie Aufbau und Abbau von virtuellen Verbindungen, Vermittlung von Paketen von Abschnitt zu Abschnitt bis zum Ziel, sowie deren Flußsteuerung.
- Ebene 4: Die Transportebene übernimmt den Nachrichtentransport zwischen den Datenendeinrichtungen über das Netz hinweg. Dazu gehören Aufgaben wie Unterteilung der Nachricht in Pakete und deren Zusammenstellung am Empfangsort, Aufbau und Beendigung von Transport-Verbindungen, Adressenzuordnung, Paketreihung und Ende-zu-Ende Flußsteuerung.
- Ebene 5: Die Sitzungsebene initiiert, überwacht und beendet Transport-Verbindungen zwischen den Anwenderprozessen. Ferner ist diese Ebene für die Synchronisation dieser Prozesse verantwortlich.
- Ebene 6: Die Darstellungsebene befaßt sich mit der Darstellung der Daten. Beispiele für Aufgaben dieser Ebene sind die Geräteanpassung, die Verschlüsselung und Entschlüsselung der Daten sowie deren Formatumsetzung.
- Ebene 7: In der Anwenderebene werden die eigentlichen Anwendungsprozesse abgewickelt. Beispiele hierfür sind Dialogbetrieb mit einem Rechner, Dateiübermittlung oder Zugriff auf Datenbanksysteme.

2.7 CCITT-Empfehlungen

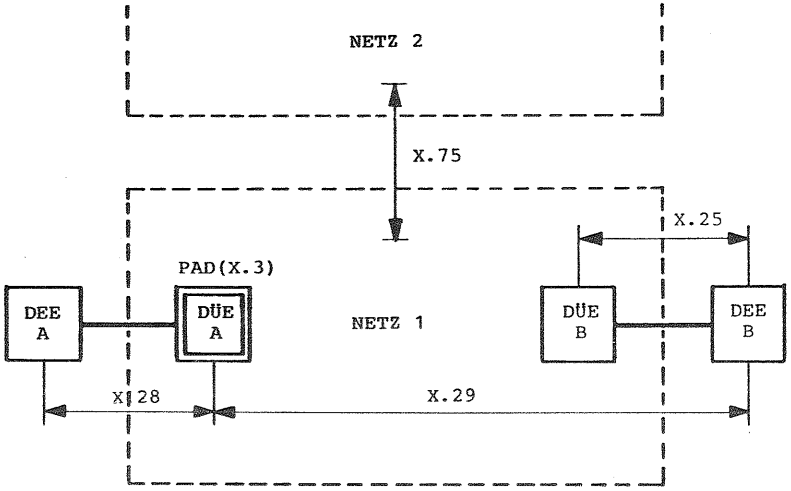
Die vom CCITT (Comité Consultatif International Télégraphique et Téléphonique) empfohlene Schnittstelle X.25 umfaßt die Ebenen 1 bis 3 des ISO-Architekturmodells. Sie definiert die geräteunabhängige Schnittstelle zwischen Datenendeinrichtung (DEE) auf der Teilnehmerseite und Datenübertragungseinrichtung (DÜE) auf der Netzseite. In dieser Empfehlung sind die wesentlichen Grundfunktionen von Paketvermittlungsnetzen festgelegt und somit kommt der Empfehlung X.25 eine zentrale Bedeutung zu.

Für zeichenorientierte Datenendeinrichtungen, die also im Start-Stop-Modus anstatt im Paketmodus arbeiten und deshalb eine

PAD-Funktion benötigen, sind die folgenden Empfehlungen maßgebend:

X.3 - **Funktionen und Parameter der PAD,**
X.28 - Schnittstelle zwischen einer Start-Stop DEE und PAD,
X.29 - Schnittstelle zwischen Paket-DEE und PAD.

Ferner gilt die Empfehlung X.75 als Grundlage für die Paketübergabe zwischen verschiedenen Paketvermittlungsnetzen. Sie beschreibt die Schnittstellen zwischen den Netzknoten. Zusammenfassend sind in Bild 2.4 die Zusammenhänge dieser CCITT-Empfehlungen [X.3, X.25, X.28, X.29, X.75] schematisch dargestellt.



DEE : Datenendeinrichtung
DÜE : Datenübertragungseinrichtung
PAD : Packet Assembly/Disassembly

DEE A : Start-Stop DEE
DEE B : Paket-DEE

Bild 2.4: Zusammenhänge zwischen wichtigen CCITT-Schnittstellen- und Protokollempfehlungen für Paketvermittlungsnetzen.

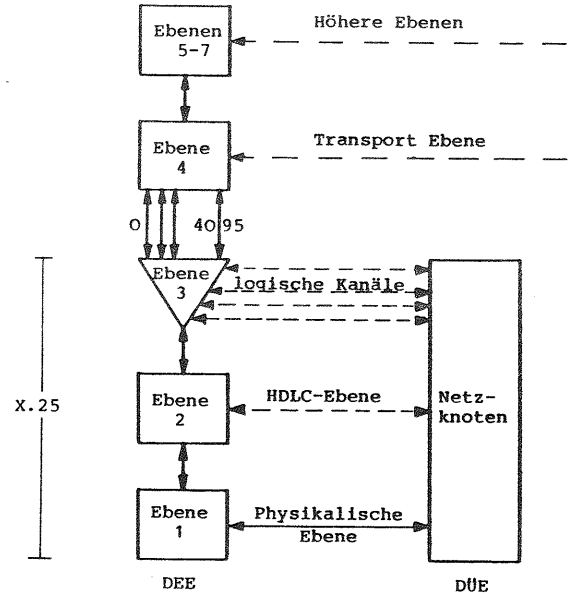


Bild 2.5: Logische Struktur der CCITT-Empfehlung X.25.

2.8 Die CCITT-Empfehlung X.25

Das Protokoll X.25 umfaßt sowohl den zeitlichen Ablauf als auch das Format der Daten, die zwischen Datenendeinrichtungen (DEE) und Datenübertragungseinrichtungen (DÜE) ausgetauscht werden. Diese für öffentliche Paketvermittlungsnetze maßgebende Schnittstelle setzt sich aus drei Ebenen zusammen. In Bild 2.5 ist die logische Struktur dargestellt. Jede Ebene übernimmt Daten der nächsthöheren Ebene, fügt sie in die vorgegebene Struktur der Steuerdaten ein und bedient sich der darunterliegenden Schicht als Kommunikationsebene. Dabei bilden Daten und Steuerungsteil der höheren Ebene die Datenfelder der nächsten Kommunikationsschicht. Dieses Ineinanderfügen der X.25-Ebenen veranschaulicht Bild 2.6.

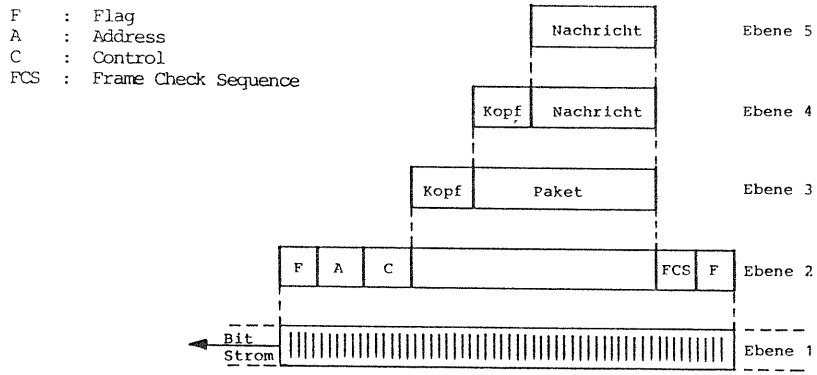


Bild 2.6: Ineinanderfügen von Datenblöcken der verschiedenen Ebenen.

An einer Schnittstelle zwischen DEE und DÜE können bis zu 15 Kanalgruppen und 255 Kanalnummern gebildet werden. Für eine virtuelle Verbindung zwischen zwei Datenendeinrichtungen ist an beiden Netzschnittstellen jeweils ein logischer Kanal notwendig. Der Datentransport im Vermittlungsnetz ist nicht näher festgelegt. Der Zusammenhang von logischen Kanälen und virtueller Verbindung zeigt Bild 2.7.

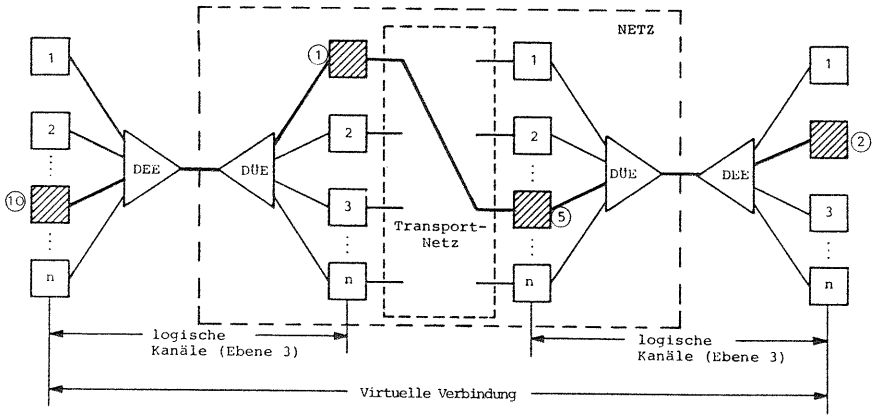


Bild 2.7: Zusammenhang zwischen logischen Kanälen und virtuellen Verbindungen.
Beispiel: virtuelle Verbindung über logische Kanäle 10 - 1 - Transportnetz - 5 - 2 .

2.8.1 Ebene 1 - physikalische Ebene

Die elektrischen, mechanischen, funktionalen und prozeduralen Parameter für die physikalische Schnittstelle sind in der CCITT Empfehlung X.21 festgelegt [X.21]. Als Funktion steht die Aufrechterhaltung der physikalischen Verbindung zwischen DEE und DÜE im Vordergrund.

Im einzelnen bedeutet dies:

- Parallel/Seriell-Wandlung,
- Anpassung an Eigenschaften der Übertragungsmedien,
- Synchronisation von Informationsbits,
- Zustandsüberwachung und -signalisation.

2.8.2 Ebene 2 - HDLC-Ebene

Die Ebene 2 legt die Prozeduren für die Übermittlung von Datenblöcken fest. Der sogenannten Version LAP B (Link Access Procedure B) liegt die HDLC-ABM-Prozedur (High Level Data Link Control, Asynchronous Balanced Mode) zugrunde. Die Hauptaufgabe der Ebene 2 besteht darin, der nächsthöheren Paketebene eine fehlerfreie Verbindung zu garantieren, wofür folgende Funktionen notwendig sind:

- Auf- und Abbau einer Verbindung zwischen DEE und Netzknoten über einen Übertragungsabschnitt,
- Steuerung des Übertragungsabschnittes für Duplexverkehr,
- Rahmenbildung und -erkennung mit Hilfe eines festgelegten Rahmenformats (HDLC-Rahmen mit Flag, Adress-, Kontroll-, Daten- und Fehlersicherungs-Feld),
- Gewährleistung für volle Datentransparenz durch Einsetzen von Null-Bits,
- Fehlererkennung durch Blocksicherungsverfahren sowie Datenflußsteuerung und Reihung mit Hilfe von fortlaufender Numerierung der Datenrahmen,
- Fehlerkorrektur von fehlerhaften Blöcken durch automatische Wiederholung,

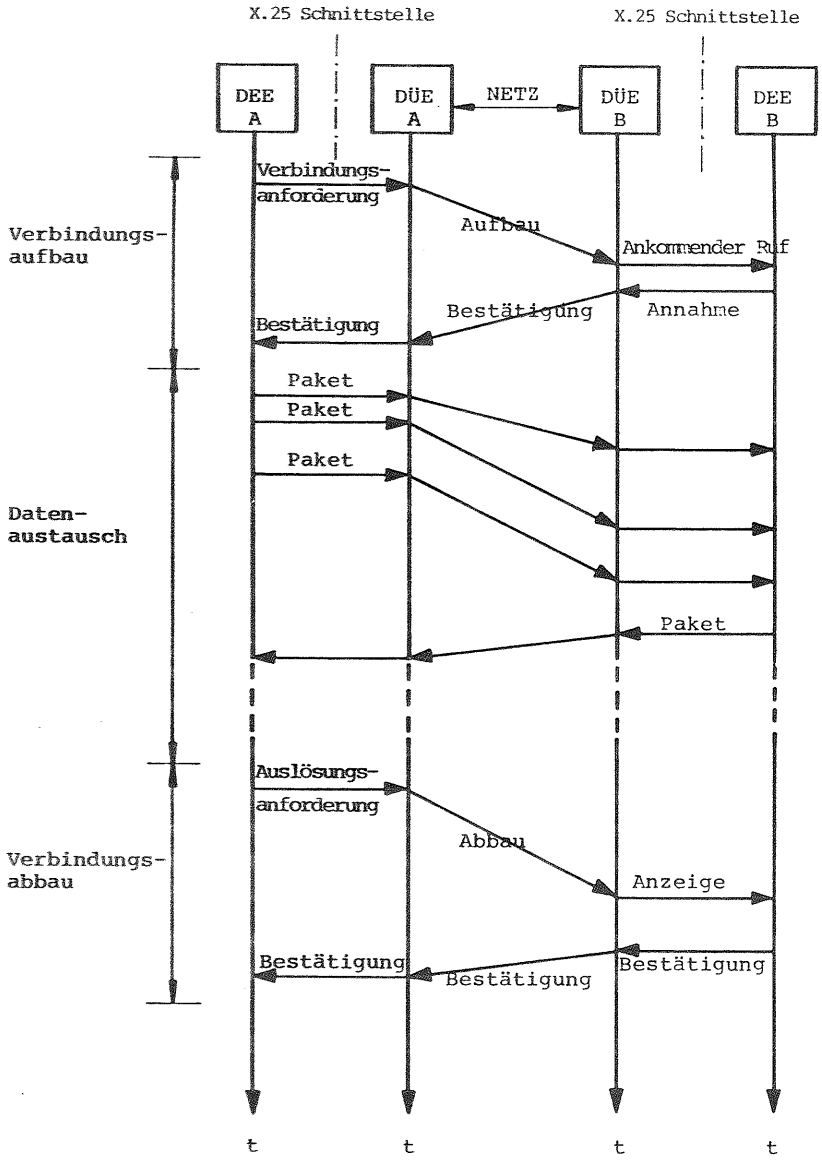


Bild 2.8: Verbindungsphasen einer virtuellen Verbindung: Aufbau, Datenaustausch, Abbau.

- Weitermelden von nicht korrigierbaren Fehlern und Protokollfehlern zur nächsthöheren Ebene.

Zur Ausführung dieser Funktion stehen eine Reihe von Befehlen und Meldungen zur Verfügung.

2.8.3 Ebene 3 - Paketebene

In dieser Ebene sind die Prozeduren festgelegt, die den Auf- und Abbau der virtuellen Verbindungen, den Datenaustausch innerhalb von virtuellen Verbindungen sowie die Einzelpaketübergabe nach dem Datagramm-Prinzip regeln. Die Eigenschaften der Ebene 3 lassen sich wie folgt zusammenfassen:

- Auf- und Abbau von virtuellen Verbindungen,
- Bereitstellung von festen virtuellen Verbindungen,
- Verwaltung der logischen Kanäle,
- Datenflußsteuerung getrennt für jede virtuelle Verbindung zwischen DEE und DÜE,
- Sicherstellung der korrekten Reihenfolge der Pakete beim Empfänger mit Hilfe einer fortlaufenden Numerierung,
- Möglichkeit zur Unterbrechung des normalen Datenflusses,
- Bereitstellung von wahlfreien Leistungsmerkmalen für den Benutzer (z.B. Mehrfachanschluß, Geschlossene Benutzergruppen, Gebührenübernahme beim Empfänger, Wahl der Durchsatzklasse bzw. Parameter der Flußsteuerung).

Pakete bestehen aus mindestens 3 Oktetts und haben üblicherweise eine maximale Länge von 128 Oktetts (Bytes).

2.8.4 Funktionsablauf einer virtuellen Verbindung

Virtuelle Verbindungen haben, ähnlich wie bei leitungsvermittelten Netzen, eine Phase der Verbindungsherstellung, eine Phase der Datenübertragung und eine Phase der Verbindungsauslösung. In welcher Reihenfolge die Befehle und Meldungen die X.25-Schnittstelle überqueren ist in Bild 2.8 schematisch dargestellt.

2.9 Datenflußsteuerung

Zur Fehlersicherung des Datentransports und zur Synchronisation vom Sender- und Empfangsprozeß ist eine Datenflußsteuerung unentbehrlich. Ohne Flußsteuerung würde eine schnelle Datenendeinrichtung, die mit einer langsamen Ziel-Datenendeinrichtung kommuniziert, die Netzabschnitte im Verbindungsweg verstopfen.

2.9.1 Funktionen der Datenflußsteuerung

Die Funktionen der Datenflußsteuerung werden hierarchisch über verschiedene Protokollebenen verteilt [Gerla/Kleinrock (1980)]:

- Auf der Sicherungsebene dient sie dazu, die Verarbeitungsgeschwindigkeit von Sender- und Empfangseinheit an den Enden des Übertragungsabschnittes aufeinander abzustimmen. Außerdem regelt diese Funktion die ordnungsgemäße Abwicklung von Wiederholungsvorgängen bei Verfälschung der Datenblöcke durch Übertragungsstörungen oder im Falle der Abweisung an der Empfangseinheit.
- Auf der Vermittlungsebene am Rand des Vermittlungsnetzes regelt die Datenflußsteuerung den Datenpaketverkehr jeder virtuellen Verbindung zwischen DEE und DÜE in Richtung zum Netz und sorgt dafür, daß die Pakete den Zielknoten möglichst schnell wieder verlassen können.
- Auf der Vermittlungsebene zwischen Ursprungs- und Zielknoten hat die Datenflußsteuerung die Aufgabe, die Anzahl der Pakete zwischen diesen beiden Netzknoten zu begrenzen und Pakete, die verlorengingen, zu wiederholen. Ein Paketverlust tritt zum Beispiel auf beim Ausfall eines Knotens oder bei einer Überlastabwehrstrategie mit Unterdrückung von blockierten Paketen. Darüberhinaus muß bei einem Paketvermittlungsnetz mit Datagramm-Betrieb im Netzinnern und mit virtuellen Verbindungen am Netzrand, am Zielknoten eine Paketreihung vorgenommen werden.
- Auf der Vermittlungsebene zwischen einzelnen Netzknoten innerhalb des Netzes begrenzt die Datenflußsteuerung die Anzahl der Pakete (zur Zeit nicht in Normen festgelegt).

- Auf der Transportebene als höchster Ebene mit einer Datenflußsteuerung existieren Aufgaben wie Begrenzung der Pakete zwischen beiden Endprozessen und der vom Benutzer möglicherweise zusätzlich noch eingesetzter Fehlersicherung und Paketreihung. Die Datenflußsteuerung ist individuell für jede Transport-Verbindung.

An dieser Stelle sei erwähnt, daß bei den verschiedenen Netzrealisierungen nicht immer alle vorher genannten Teilaufgaben vorhanden sind.

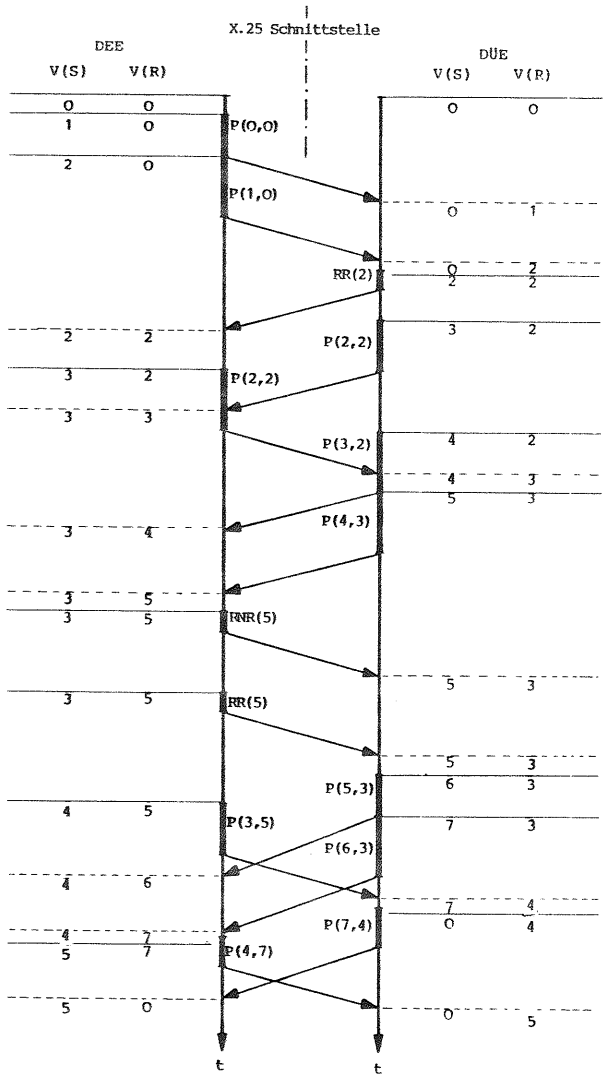
2.9.2 Fenstermechanismus

Der Fenstermechanismus ist heute eine weitverbreitete Methode der Datenflußsteuerung. Die Benennung der Steuergrößen, Befehle und Meldungen sowie die Implementierungen können zwar je nach Protokollhierarchie und Paketvermittlungsnetz verschieden sein, das Prinzip bleibt jedoch gleich.

Als Beispiel sei hier die Datenflußsteuerung von X.25-Ebene 3 betrachtet. Für die Regulierung des Paketflusses an der Schnittstelle DEE/DÜE wird für jede Richtung einer virtuellen Verbindung eine Fenstergröße W festgelegt. Das Fenster gibt an, wieviel aufeinanderfolgende Pakete in der betrachteten Richtung ohne Quittung die Schnittstelle überschreiten dürfen. Die Fenstergröße W ist im allgemeinen 2, kann aber für jede Richtung individuell anders festgelegt werden.

Für die Flußsteuerung wird an beiden Seiten der Schnittstelle ein Satz von Steuergrößen eingerichtet. Dieser besteht im wesentlichen aus einer Paket-Sendefolgevariable V(S), einer Paket-Empfangsfolgevariable V(R) und der Fenstergröße W. Ein Ausschnitt aus dem Protokoll-Ablauf ist in Bild 2.9 dargestellt.

Zu Beginn sind die Werte von V(S) und V(R) in beiden Richtungen gleich 0. Danach durchlaufen sie die zulässigen Werte, bei modulo 8 wird z.B. zyklisch von 0 bis 7 durchnumeriert. Wird ein Fenster W=2 verwendet, so dürfen maximal 2 Pakete gesendet werden, bevor eine Quittung eintrifft. Die Werte von V(S) und V(R) erlauben es festzustellen, welche Paketnummern empfangen



P(s,r) = Datenpaket mit P(S)=s und P(R)=r
RR(r) = RR -Paket mit P(R)=r
RNR(r) = RNR-Paket mit P(R)=r

DEE = Datenendeinrichtung
DÜE = Datenübertragungseinrichtung
V(S) = Paket-Sendefolgevariable
V(R) = Paket-Empfangsfolgevariable

Zustand der Variablen V(S), V(R) : vorher / nachher SENDEN vorher / nachher EMPFANG

Bild 2.9: Fenstermechanismus: Beispiel eines Ablaufes für die X.25 Ebene 3.

werden dürfen, welche zu bestätigen sind und ob ein neues Paket gesendet werden darf.

Quittungen werden entweder den Datenpaketen der Gegenrichtung beigepackt (piggybacking) oder werden als spezielle Steuerpakete zurückgesendet. Das Steuerpaket RR (Receiver Ready) wird verwendet, wenn der Empfänger weitere Pakete aufnehmen kann, das RNR-Paket (Receiver Not-Ready) wird gesendet, wenn der Paketfluß vorübergehend gestoppt werden soll.

Die Variablen V(S) und V(R) auf beiden Seiten der Schnittstelle DEE/DÜE werden fortlaufend aktualisiert. Dazu wird in jedem Datenpaket eine Sende- und eine Empfangsfolgenummer, P(S) und P(R), mitgeführt. In Steuerpaketen (RR, RNR) wird lediglich P(R) übermittelt. Der Informationsaustausch zwischen den Steuergrößen findet nach den folgenden Regeln statt:

- Datenpakete dürfen gesendet werden, wenn das Fenster "offen" ist, das heißt wenn die Bedingung $V(R) \leq P(S) < V(R) + W$ erfüllt ist. Siehe dazu Bild 2.10.
- Wird ein Datenpaket gesendet, so wird der aktuelle Stand von V(S) bzw. V(R) in die Felder P(S) bzw. P(R) des Paketkopfes geschrieben und anschließend wird V(S) um 1 erhöht (Modulo-Rechnung).
- Beim Empfang eines Datenpaketes wird zuerst die mitgeführte Sendefolgenummer P(S) mit dem Wert der Variablen V(R) verglichen. Bei Übereinstimmung wird das Datenpaket aufgenommen und V(R) um 1 erhöht, anderenfalls liegt ein Sequenzfehler vor, und das Paket wird ignoriert.
- Beim Senden von Steuerpaketen (RR, RNR) wird P(R)=V(R) gesetzt.

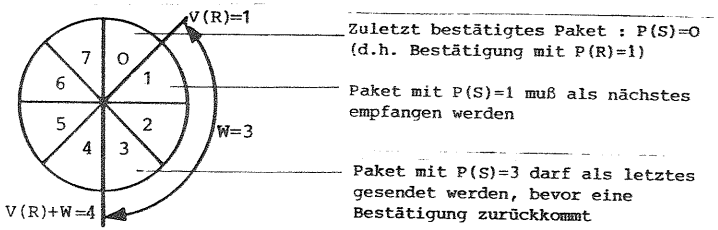


Bild 2.10: Erläuterung zur Fenstermechanismus

2.10 Automatische Fehlerkorrektur-Verfahren

Erkennung und Behandlung von fehlerhaften Paketen stehen in engem Zusammenhang mit der Datenflußsteuerung. Auch hier sind die entsprechenden Aufgaben hierarchisch über die einzelnen Protokollebenen verteilt.

In einem Paketvermittlungsnetz erfordern die folgenden Situationen eine Fehlerbehandlung:

- Bitübertragungsfehler durch Störungen auf der Übertragungsleitung, Hardware-Defekte oder ähnliche Ursachen,
- fehlende Pakete durch Ausfall eines Netzknotens oder einer Leitung, Softwarefehler oder die bewußte Unterdrückung von Paketen in Überlastsituationen,
- Pakete mit Sequenzfolgefehlern durch Verlust eines Paketes oder bei individueller Übermittlung von Paketen über unterschiedliche Übertragungswege,
- Paketkopien durch Wiederholung eines Paketes bei Verlust oder verzögertem Eintreffen der Quittung (Time-Out).

2.10.1 Abschnittsweise Fehlersicherung

Zur Erkennung und Behebung von Übertragungsfehlern enthält jeder Datenblock einen Fehlersicherungsteil, dessen Kontrollbits sich aus den Informationsbits nach einem bestimmten Schema ableiten lassen. Diese Fehlersicherung findet abschnittsweise statt und wird von der Sicherungsebene (Ebene 2) ausgeführt. Die Fehlererkennung kann entweder eine Paritätsprüfung sein oder wird als zyklische Redundanzprüfung mit Hilfe von Generatorpolynomen durchgeführt [Swoboda (1973), Tanenbaum (1981)].

Beim Empfänger werden die Kontrollbits nach demselben Schema wie beim Sender ermittelt, so daß sich mit einem Vergleich zwischen errechneten und empfangenen Kontrollbits fehlerhafte Blöcke erkennen lassen. Blöcke mit Übertragungsfehlern werden in Paketvermittlungsnetzen durch eine wiederholte Übertragung korrigiert (ARQ-Methode, Automatic Repeat Request).

Bei dieser Methode wird beim Sender eine Kopie des übertragenen Blockes gespeichert. Der Empfänger prüft den empfangenen Block auf Fehlerfreiheit. Bei einem positiven Ergebnis nimmt der Empfänger den Block an und es wird eine positive Quittung zurückgesendet, so daß die gespeicherte Kopie gelöscht werden kann. Ein fehlerhafter Block wird vom Empfänger unterdrückt und dessen Wiederholung wird mit einer negativen Quittung veranlaßt.

Als Antwort auf eine negative Quittung stehen zwei Wiederholungsmechanismen zur Auswahl:

- Bei nichtselektiver Wiederholung werden sowohl der fehlerhafte Block als auch alle nachfolgenden Blöcke wiederholt. Am Empfänger werden alle Blöcke bis zum Eintreffen des angeforderten Blockes unterdrückt. Dadurch wird stets die richtige Reihenfolge der Blöcke beim Empfänger gewährleistet.
- Bei selektiver Wiederholung wird nur der fehlerhafte Block wiederholt. In diesem Falle muß der Empfänger selbst die richtige Reihenfolge der empfangenen Blöcke wiederherstellen. Dazu ist neben dem erhöhten Aufwand an Intelligenz auch zusätzlicher Speicheraufwand erforderlich.

Zusätzlich wird der letzte Block in einer Folge von Blöcken zeitüberwacht. Dazu wird eine Zeitbegrenzung bei Beginn der Blockübertragung gesetzt und beim Eintreffen der entsprechenden Quittung zurückgesetzt. Bei Ablauf der Zeitüberwachung (Time-Out) wird dieser letzte Block erneut gesendet oder es werden Maßnahmen getroffen, um den Grund für das Ausbleiben der letzten Quittung feststellen zu können. Die Anzahl der Wiederholungen ist begrenzt.

2.10.2 Ende-zu-Ende Fehlersicherung

Währenddem die Bitübertragungsfehler abschnittsweise erkannt und korrigiert werden, ist es die Aufgabe der Ende-zu-Ende Fehlersicherung, fehlende Pakete, doppelte Pakete sowie Sequenzfolgefehler zu erkennen und anschließend zu beheben. In Paketvermittlungsnetzen, in denen die Pakete einer virtuellen Verbindung über verschiedene Wege vermittelt werden dürfen, führt sie darüberhinaus die Paketreihung durch.

Die Erkennung am Empfangsort erfolgt mit Hilfe der Sequenzfolgenummern, die am Sendeort durch die Überwachung der Paketquittierungszeit (Time-Out).

Beim Empfang eines doppelten Paketes wird dieses Paket lediglich unterdrückt, in allen anderen Fällen geschieht die Fehlerbehebung durch Paketwiederholung. Paketwiederholungen werden entweder anläßlich diesbezüglicher Anforderungen (selektiv oder nicht selektiv) oder infolge einer Quittierungszeitüberschreitung vorgenommen. Zu diesem Zweck muß für jedes Paket bis zur Quittierung eine Kopie am Sendeort abgespeichert werden.

2.11 Verkehrslenkung

Verkehrslenkungsalgorithmen werden gemäß dem ISO-Architekturmodell in der Ebene 3 der Netzknoten implementiert. Hier gilt es zu entscheiden, über welche Ausgangsleitung ein eintreffendes Paket weitervermittelt werden soll. Die Verkehrslenkung wird im Datagramm-Betrieb für jedes Paket individuell vorgenommen. In einem Netz mit virtuellen Verbindungen jedoch werden die Entscheidungen für die Verkehrslenkung nur beim Aufbau der Verbindung durchgeführt, danach werden alle Pakete über den gleichen Weg übermittelt.

In beiden Fällen werden mit Hilfe der Verkehrslenkung möglichst kurze Netzdurchlaufzeiten für die Pakete sowie ein guter Belastungsausgleich im Netz angestrebt. Dabei ist es wesentlich, daß der Verkehrslenkungsalgorithmus stabil und unempfindlich ist gegen Änderungen in der Netztopologie.

Im wesentlichen wird unterschieden zwischen:

- fester Verkehrslenkung (fixed routing) mit fest vorgeschriebenen Wegen und ohne die Möglichkeit zur Anpassung an die Verkehrsverhältnisse. Bei Ausfällen von Netzkomponenten können die Pakete oft auf Alternativwegen, die ebenfalls vorher festgelegt sind, umgeleitet werden. Vorteile sind eine einfache Handhabung der Verkehrslenkungstabellen und der geringe Verwaltungsaufwand.

Die Initialisierung der Verkehrslenkungstabellen basiert entweder auf Methoden des kürzesten Weges (shortest path or minimum cost routing), wobei die Kostenfunktion von der Netzimplementierung abhängt, oder auf Methoden der geringsten Zeitverzögerung (minimum average delay routing), bei der die erwarteten Verkehrsbeziehungen mitberücksichtigt werden.

- adaptiver Verkehrslenkung (adaptive routing) mit Verkehrslenkungstabellen, die ständig an die neuen Verkehrsverhältnisse angepaßt werden. Die Aktualisierung der Tabellen ist jedoch ein beträchtliches Problem, und der Entscheidungsprozeß für den günstigsten Weg basiert oft auf veralterten Informationen. Es werden lokale Zustandsinformationen, Informationen von benachbarten Netzknoten und Informationen von einem Netzkontrollzentrum herangezogen. Angewendet werden Algorithmen wie die Methode der kürzesten Warteschlangenlänge (shortest queue routing), die Methode der minimalen Verzögerung (minimum estimated delay routing) und Delta-Verkehrslenkung (Delta-routing).

Die Vielfalt der Verkehrslenkungsmethoden ist z.B. beschrieben in [Rudin (1976), Greene/Pooch (1977), Schwartz (1977), Davies/Barber/Price/Solomonides (1979), Schwartz/Stern (1980), Tanenbaum (1981)].

3. ÜBERLASTPROBLEMATIK IN PAKETVERMITTLUNGSNETZEN

Das Verkehrsgeschehen in Paketvermittlungsnetzen, und somit auch die Überlastproblematik, wird bestimmt durch die Wechselwirkung zwischen dem stochastischen Verkehr und der Betriebsmittelvergabe. Das Problem der Überlast und die Abwehrmaßnahmen umfaßt deshalb:

- Charakterisierung des angebotenen Verkehrs,
- Aspekte zur verkehrsgerechten Realisierung von Paketvermittlungsnetzen,
- Ursachen und Indikatoren für Überlastsituationen,
- Klassifizierung und Beschreibung von Überlastabwehrstrategien.

3.1 Charakterisierung des angebotenen Verkehrs

Die Verkehrscharakteristiken in Paketvermittlungsnetzen sind durch folgende zwei grundsätzlichen Betriebsarten bestimmt:

a) Dialogbetrieb mit den Merkmalen:
   - niedriges Datenvolumen,
   - vorwiegend kurze Pakete,
   - kurze Übertragungsdauer und lange Pausen,
   - kurze Quittierungszeiten erforderlich,
   - Hauptverkehr tagsüber.

b) Stapelbetrieb mit den Merkmalen:
   - hohes Datenvolumen,
   - vorwiegend Pakete mit maximaler Länge,
   - lange Übertragungsdauer und kurze Pausen,
   - längere Quittierungszeiten zugelassen,
   - Hauptverkehr häufig erst abends und nachts.

Darüberhinaus unterscheidet man zwischen den Anforderungen für den Verbindungsaufbau und dem Datenaustausch selbst. Abgesehen von sporadischen Häufungen spielt der Verkehrsanteil der Verbindungsaufbau-Pakete eine untergeordnete Rolle. Das gesamte

Verkehrsangebot ist jedoch abhängig von der Anzahl der gleichzeitig bestehenden virtuellen Verbindungen und dem stochastischen Ablaufgeschehen innerhalb dieser Verbindungen.

3.2 Aspekte zur verkehrsgerechten Realisierung von Paketvermittlungsnetzen

Die Realisierung von Paketvermittlungsnetzen beruht auf dem Zusammenspiel von Hardware- und Softwaretechnologie, Übertragungs- und Vermittlungstechnik, Wirtschaftlichkeit und verkehrstheoretischen Gesichtspunkten.

Dabei ist die Verwirklichung der einzelnen Verkehrskriterien bedingt durch die Kapazität der Übertragungsstrecken, die Rechnerleistung sowie die Speicherkapazität der Netzknoten und die Regeln, nach denen diese Betriebsmittel den Paketen zugewiesen werden (z.B. Speicherzuteilung, Verkehrslenkung, Datenflußsteuerung und Überlastabwehrstrategie).

Entwurf, Dimensionierung und Betrieb von Paketvermittlungsnetzen werden somit von den folgenden, teilweise widersprüchlichen Zielsetzungen entscheidend beeinflußt:

- niedrige Quittierungszeiten für Dialogverkehr und Realzeitanwendungen,
- hoher Durchsatz für Stapelverkehr,
- faire Zuteilung der Betriebsmittel an die einzelnen Anwenderprozesse,
- optimale Auslastung der Betriebsmittel,
- hohe Netzverfügbarkeit,
- allgemeine Zugänglichkeit des Netzes,
- hohe Zuverlässigkeit der Paketübermittlung,
- und vor allem Wirtschaftlichkeit.

Die Quittierungszeit, definiert als die Zeitdauer zwischen dem Absenden eines Paketes und der Bestätigung über den korrekten Empfang am Zielort, ist für den Dialog- und Realzeitbetrieb dann optimal ausgelegt, wenn sie mit einer gewissen Wahrschein-

lichkeit eine Zeitobergrenze nicht überschreitet (z.B. 95 % weniger als 0.5 Sekunden).

Kurze Quittierungszeiten bedingen im wesentlichen eine geringe Anzahl von Übertragungsabschnitten und möglichst kurze Warteschlangenlängen in den einzelnen Netzknoten. Auch eine bevorzugte Abfertigung kann die Wartezeiten für dringende Pakete in den Netzknoten verringern.

Der Durchsatz, definiert als die mittlere Anzahl von Paketen pro Sekunde, steht dagegen bei Stapelbetrieb im Vordergrund. Für einen hohen Durchsatz soll die Datenflußsteuerung zwischen Ursprung und Ziel den Datenaustausch möglichst wenig bremsen. Dies verlangt ein großes Fenster. Somit können große Warteschlangen entstehen und die Quittierungszeiten sind entsprechend länger.

Eine faire Zuteilung der Betriebsmittel erfordert Regeln, um zu verhindern, daß Anwenderprozesse aufgrund einer günstigen geographischen Position oder infolge eines hohen Datenvolumens in der Lage sind, einen übermäßigen Betriebsmittelanteil zu belegen.

Eine optimale Auslastung der Betriebsmittel steht in engem Zusammenhang mit den drei vorangehenden Kritierien; eine entsprechende Optimierung ist abhängig von der Gewichtung dieser Zielsetzungen.

Eine hohe Netzverfügbarkeit erfordert eine Redundanz vieler Hardwarekomponenten in den Netzknoten und die Möglichkeit, bei Bedarf auf andere Übertragungsstrecken auszuweichen.

Eine allgemeine Zugänglichkeit bedingt eine Vielzahl von Zusatzeinrichtungen, die es erlauben, zeichenorientierte Datenendgeräte am Paketvermittlungsnetz anzuschließen oder einen Zugang über andere Nachrichtennetze zu ermöglichen.

Eine hohe Zuverlässigkeit der Paketvermittlung fordert neben der Fehlersicherung auf den einzelnen Übertragungsstrecken das Aufbewahren von Paketkopien bis zur Bestätigung des korrekten

Empfangs und die Paketwiederholung beim Ausbleiben einer Quittung. Die Fehlersicherung auf den verschiedenen Protokollebenen verlängert jedoch auch die Quittierungszeit, und vorzeitige Paketwiederholungen wirken durchsatzmindernd.

Wirtschaftlichkeit verhindert schließlich eine Überdimensionierung der Betriebsmittel. Dabei gilt: je wirtschaftlicher das Netz ausgelegt ist, desto weniger Spielraum verbleibt für die Bewältigung eines erhöhten Verkehrsangebots.

3.3 Überlastsituationen

In Analogie zu dem Straßenverkehr ist das Entstehen von Überlastsituationen in Paketvermittlungsnetzen von komplexer Natur. Solche Situationen sind typisch für die dynamische Wechselwirkung zwischen stochastischem Verkehrsangebot und einem verteilten System mit einer beschränkten Zahl von Betriebsmitteln.

Kennzeichnend für ein überlastetes Netz oder Netzteil ist eine Abnahme des Durchsatzes und eine Verlängerung der Quittierungszeiten. Das typische Verhalten ist in Bild 3.1 veranschaulicht.

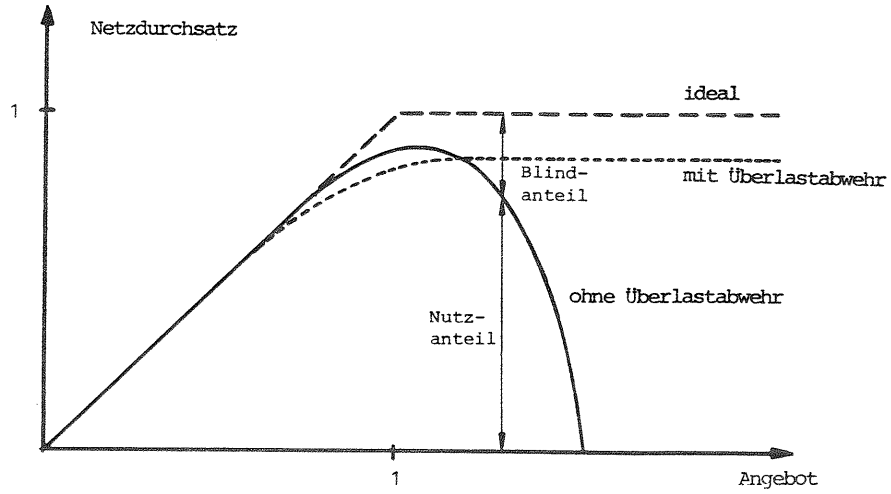


Bild 3.1: Charakteristik eines überlasteten Paketvermittlungsnetzes.

Solange das Verkehrsangebot unterhalb der Verkehrskapazität des betrachteten Netzteiles bleibt, steigt der Durchsatz mit wachsendem Verkehrsangebot an: zuerst linear und anschließend etwas geringer. Wird jedoch in einem Netz ohne Überlastabwehrstrategie die Verkehrskapazität überschritten, so nimmt der Blindlastanteil durch Paketwiederholungen und gegenseitige Verkehrsbehinderung sehr stark zu, so daß der Durchsatz entsprechend geringer wird. Im Extremfall kann eine Verklemmungssituation auftreten.

Dieses ungünstige Verhalten kann mit Hilfe von verschiedenen Abwehrmaßnahmen verhindert werden.

3.3.1 Ursachen

Verschiedene Ursachen können eine Überlast auslösen:

- Unangemessene System- und Netzplanung können Systemengpässe verursachen, die sich oft auch auf andere Teile des Netzes auswirken. Zu diesen systeminternen Ursachen gehören ungenügende Speicher-, Verarbeitungs- und Übertragungskapazitäten, ungünstige Ablaufsteuerung in den Netzknoten, Vermittlungsprotokolle.
- Der Ausfall von Übertragungsabschnitten, Verarbeitungs- oder Speichereinheiten bedeutet eine Kapazitätsminderung und je nach Systemrekonfigurationsverhalten kann dies eine Überlastsituation auslösen.
- Eine große Anzahl von gleichzeitig bestehenden virtuellen Verbindungen, die alle für sich eine Datenflußsteuerung besitzen, können wegen ihrer stoßartigen Betriebsweise temporär ein Gesamtverkehrsangebot erzeugen, das ein Vielfaches der Nennkapazität der Netzkomponenten beträgt. Es entstehen also ausgeprägte Lastspitzen.
- Eine temporäre Blockierung am Empfangsort verursacht einen Rückstau von Paketen im Netz, der eine Überlastsituation verursachen kann.
- Ein langandauernder hoher Datenverkehr kann ebenfalls eine Überlastsituation entstehen lassen.

- Eine Häufung von Anforderungen für den Aufbau von virtuellen Verbindungen kann morgens auftreten, wenn Banken, Geschäfte und Reisebüros Verbindungen mit ihren Kommunikationspartnern initiieren. Ein derart hoher Verbindungsaufbauaufwand kann die betroffenen Rechner in den Netzknoten stark überlasten.

Als Reaktion auf diese Primärursachen können einige Nebenwirkungen entstehen, die die Überlastsituation noch verschärfen:

- Paketwiederholungen verursachen eine zusätzliche Netzbelastung, die im Normalfall zu vernachlässigen ist. Werden jedoch Paketwiederholungen als Reaktion auf eine Überlastsituation im Netz ausgelöst, so kann diese Blindlast ein beträchtliches Ausmaß annehmen und das Netz derart überlasten, daß der Paketverkehr total zusammenbricht.
- Überlastete Verkehrsströme beanspruchen infolge blockierter Pakete einen hohen Speicherbedarf. Durch diesen Speicherengpaß werden auch alle anderen Verkehrsströme behindert und dies kann leicht dazu führen, daß die Übertragungskapazität nicht voll ausgenutzt werden kann.
- Insbesondere in Überlastsituationen können als Folge von gegenseitigem Warten auf das Freiwerden von Betriebsmitteln Verklemmungssituationen (Deadlocks) entstehen.

3.3.2 Überlastindikatoren

Zur möglichst frühzeitigen Erkennung von Überlastsituationen benötigt ein Netzknoten Indikatoren.

Wie dies in Bild 3.2 dargestellt ist, werden diese Überlastindikatoren ermittelt aus:

- Messungen im Netz und im Netzknoten selbst,
- internen Zustandsinformationen,
- Meldungen von anderen Netzknoten und vom Netzkontrollzentrum.

Die Durchführung einer Messung (Zeiten oder Zählvorgänge) erfordert ein entsprechendes Meßintervall, so daß die Meßwerte einen Zustand in der unmittelbaren Vergangenheit darstellen. Die Präzision, mit der das zukünftige Netz- oder Systemver-

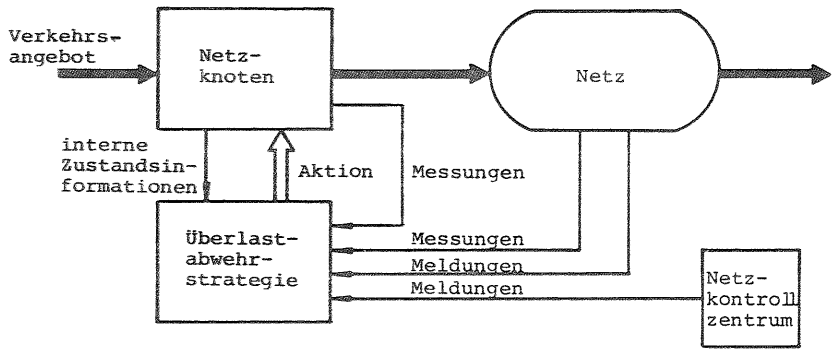


Bild 3.2: Ermittlung von Überlastindikatoren.

halten vorausgesagt werden kann, hängt vom Schätzungsverfahren, von der Häufigkeit und der Genauigkeit der Messungen sowie von der Länge des Meßintervalls ab.

Zu dieser Kategorie von Überlastindikatoren gehören:

- Auslastung der Übertragungsstrecken, aus der die gesamte Verkehrslast (Pakete, Wiederholungen und Verwaltungsarbeit des Übertragungsprotokolls) auf den einzelnen abgehenden Übertragungsstrecken ermittelt werden kann.
- Anzahl der Paketwiederholungen (Vermittlungsebene) als Indikator für möglicherweise überlastete Zielrichtungen.
- Quittierungsverzögerung als Maß für die Quittierungszeiten der verschiedenen Zielrichtungen.

Bei internen Zustandsinformationen handelt es sich um einen gegenwärtigen Zustand, aus dem direkt auf eine bevorstehende Überlastsituation rückgeschlossen werden kann.

Dies geschieht durch die Überwachung von:

- Warteschlangenlängen, aus denen einerseits die zukünftige Netzlast in die verschiedenen Zielrichtungen ermittelt werden kann und die andererseits als Maß für einen Rückstau dienen.
- Speicherbelastung (Pakete und Kopien) zur Anzeige von drohenden Engpässen.
- Anzahl der gesetzten Time-Outs (Vermittlungsebene), aus der sich die im Netzknoten noch zu quittierende Netzlast je Richtung ermitteln läßt.

Durch Meldungen von benachbarten Netzknoten oder von den Zielknoten wird jeder Netzknoten über deren Belastungssituation in Kenntnis gesetzt. Diese Meldungen erfolgen entweder in regelmäßigen Abständen oder aufgrund einer wesentlichen Belastungsänderung.

Darüber hinaus erhält jeder Netzknoten Meldungen vom Netzkontrollzentrum.

3.4 Klassifizierung und Beschreibung von Überlastabwehrstrategien

Zur Verhinderung von Überlastsituationen, die insbesondere in Paketvermittlungsnetzen völlig unerwartet entstehen können, werden neben Methoden für die Erkennung von Überlast diverse Abwehrstrategien benötigt.

Dabei sind bei der Entwicklung und Implementierung von Überlastabwehrstrategien folgende Ziele zu verfolgen:

- Die Wirkungsgeschwindigkeit soll dem geographischen Wirkungsbereich (lokal oder global) angepaßt sein.
- Anstatt einer abrupten soll eine gleichmäßige Wirkung erzielt werden.
- Der Verwaltungsaufwand soll niedrig und unabhängig von der Netzbelastung sein.
- Die Maßnahmen sollen selektiv sein, so daß sie sich lediglich auf bestimmte Verkehrsströme auswirken.
- Die Maßnahmen sollen jedoch auch fair sein, so daß keine Verkehrsströme übermäßig benachteiligt werden.

3.4.1 Hierarchische Gliederung

Eine effiziente Überlastabwehr in Paketvermittlungsnetzen erfordert eine Hierarchie von genau aufeinander abgestimmten Abwehrmaßnahmen. Bild 3.3 zeigt die hierarchische Gliederung der Überlastabwehrstrategien:

- Auf der Vermittlungsebene existieren drei Bereiche, und zwar der Netzzugang, der Bereich zwischen Ursprungs- und Zielknoten und der Bereich zwischen benachbarten Netzknoten.
- Auf der Transportebene befindet sich die Datenflußsteuerung zwischen den Anwenderprozessen, die für diesen individuellen Verkehrsfluß Überlastabwehrfunktionen zu erfüllen hat.
- Auf der hierarchisch höchsten Ebene ist das Netzkontrollzentrum zu finden.

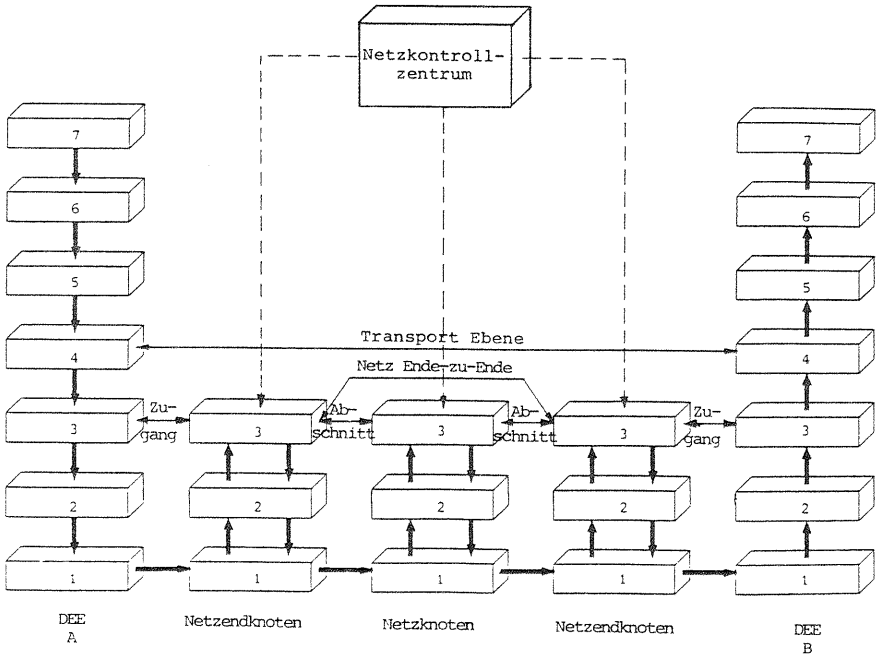


Bild 3.3: Hierarchische Gliederung von Überlastabwehrstrategien.

3.4.2 Wirkungsbereich und Wirkungsgeschwindigkeit

In engem Zusammenhang mit der hierarchischen Implementierung stehen auch der geographische Wirkungsbereich und die Wirkungsgeschwindigkeit der Überlastabwehrstrategien.

In bezug auf den geographischen Wirkungsbereich kann unterschieden werden zwischen:

- lokalen Überlastabwehrstrategien: Netzzugang, Netzknoten selbst. Sie basieren auf Informationen, die im Netzknoten selbst aufgrund seines internen Zustandes, aufgrund von Messungen oder Meldungen vorhanden sind. Die lokale Überlastabwehrstrategie soll dem Entstehen von lokalen Überlastsituationen oder Verkehrsbehinderungen zuvorkommen.
- globalen Überlastabwehrstrategien: gesamtes Netz, Netzbereich, benachbarte Netzknoten, Beziehungen zwischen Netzendknoten, Beziehungen zwischen Anwenderprozessen, Regelung der Netzzugänge. Sie beruhen auf einer statisch oder dynamisch festgelegten Begrenzung der Anzahl von Paketen und sie sollen vermeiden, daß diese Zahl überschritten wird. Die zur Überlastabwehrmaßnahme notwendigen Informationen stehen i.a. nicht sofort zur Verfügung, sondern müssen durch Meldungen von entfernten Netzknoten bereitgestellt werden.

Da einerseits eine globale Überlastabwehrstrategie die Bildung von lokalen Überlastsituationen in ihrem Wirkungsbereich nicht verhindern kann und andererseits eine lokale Überlastabwehrstrategie nur eine suboptimale Auslastung der einzelnen Betriebsmittel bewirken kann, ist für einen optimalen Netzbetrieb ihre Kombination unentbehrlich. Die lokal wirkenden Überlaststrategien werden in den Kapiteln 6 und 7, die global wirkenden in Kapiteln 8 und 9 näher betrachtet.

Hinsichtlich der Wirkungsgeschwindigkeit gilt allgemein, daß je höher die hierarchische Implementierungsstufe ist, und damit je größer der geographische Wirkungsbereich, desto langsamer kann und darf die Überlastabwehrstrategie wirken.

Damit gilt für die Rangordnung in abnehmender Wirkungsgeschwindigkeit: Netzknoten selbst - Umgebung der Netzknoten - Netz-

bereich bzw. Beziehung zwischen Netzendknoten oder Anwenderprozessen - gesamtes Netz. Der Charakter der Überlastabwehr ändert sich somit vom hochdynamischen zum quasi-statischen.

Über die Zeitspanne zwischen dem Moment der Überlasterkennung und dem Beginn der Abwehrmaßnahmen können folgende Aussagen gemacht werden:

- Für Überlastabwehrstrategien, die den Netzknoten selbst betreffen, kann die entsprechende Abwehrmaßnahme sofort getroffen werden, da die Erkennung und die Ausführung direkt im Netzknoten stattfindet. Es handelt sich hierbei um die Umorganisation der Paketspeicherung, Änderung in der Abfertigungsreihenfolge oder um eine Umstellung auf eine andere Paketannahmestrategie.
- Überlastabwehrstrategien, die den Netzzugang regeln, wirken ebenfalls verzögerungsfrei, sofern die Überlasterkennung im Netzendknoten selbst stattfindet. Werden die Abwehrmaßnahmen von einem anderen Netzknoten eingeleitet, z.B. mit Drosselungspaketen, so muß eine entsprechende Verzögerungszeit für das Eintreffen dieser Pakete berücksichtigt werden.
- Bei Überlastabwehrstrategien, die sich über das Paketvermittlungsnetz hinweg oder einen Teil des Netzes erstrecken, muß eine stochastische Wirkungsverzögerung einbezogen werden. Dies ist auch dann notwendig, wenn diese Netzsteuerungspakete mit Priorität vermittelt werden. Die Größe der Verzögerung hängt im wesentlichen von der geographischen Distanz ab.

3.4.3 Art der Realisierung

Bei einer Klassifizierung der Überlastabwehrstrategien nach der Art ihrer Realisierung können die folgenden Möglichkeiten identifiziert werden:

- Speicherverwaltung,
- Datenflußsteuerung,
- Ablaufsteuerung,
- Verkehrslenkung,
- Topologie,
- Tarifgestaltung.

3.4.3.1 Speicherverwaltung

Eine umfangreiche Gruppe von Überlastabwehrstrategien wird mit Hilfe der Speicherverwaltung realisiert. Im wesentlichen geht es hier um Regeln zur gezielten Abweisung von Paketen. Das Ziel ist die Gesamtoptimierung der Netzbetriebsmittel unter beliebigen Lastsituationen. Da es sich hierbei um lokale Überlastabwehrstrategien handelt, kann dieses Ziel nur zum Teil erreicht werden.

Bei völliger Speicherreservierung findet nie ein Speicherüberlauf statt. Die Speicher sind dann aber äußerst gering ausgenutzt. Werden hingegen die Speicherplätze völlig nach dem momentanen Bedarf vergeben, so müssen Pakete bei einem vollen Speicher willkürlich abgewiesen werden, so daß durch Paketwiederholungen eine erhebliche Blindlast entstehen kann.

Überlastabwehrstrategien verwenden deshalb eine Speicherreservierung, die zwischen diesen beiden Extremfällen liegt. Ihre Realisierungsform hängt von den Kriterien zur Unterscheidung der Pakete ab.

Einige Kriterien hierzu sind:

a) Unterscheidung der Pakete nach abgehenden Richtungen

Maßgebend hierfür ist die Speicherorganisation für die einzelnen abgehenden Richtungen, die in folgende Software-Organisationsformen unterteilt werden können:

- völlig gemeinsamer Speicherbereich (CS, complete sharing),
- völlig getrennte Speicherbereiche (CP, complete partitioning),
- gemeinsamer Speicherbereich mit richtungsabhängiger Begrenzung (SMXQ, sharing with maximum queue length),
- gemeinsamer Speicherbereich mit richtungsabhängiger Reservierung (SMA, sharing with minimum allocation),
- gemeinsamer Speicherbereich mit richtungsabhängiger Begrenzung sowie Reservierung (SMQMA, sharing with maximum queue length and minimum allocation).

Bei einem völlig gemeinsamen Speicher für sämtliche Richtungen (CS) ist der Speicher vorwiegend mit Paketen eines dominierenden Verkehrsstromes belegt. Pakete, die für eine Richtung bestimmt sind und einen kleineren Verkehr darstellen, müssen deshalb öfters wegen eines verstopften gemeinsamen Speichers abgewiesen werden, obwohl der betreffende Übertragungskanal größtenteils frei ist. Andererseits ist bei völlig getrennten Speicherbereichen (CP) die Speicherausnutzung gering.

Diese beiden Organisationsformen sind deshalb nicht sehr geeignet. Durch eine richtungsabhängige Begrenzung der maximalen Speicherplatzbelegung (SMXQ) kann vermieden werden, daß Verkehrsströme den Gesamtspeicher verstopfen können. Es stellt sich heraus [Irland (1978)], daß der optimale Wert für die Speicherbegrenzung - einheitlich für alle Richtungen - eine komplizierte Funktion der einzelnen Verkehrswerte ist. Bei jeder Änderung in den Verkehrsverhältnissen muß dieser Wert somit neu festgelegt werden.

Mit einer heuristischen Approximation, die als die "Quadratwurzelstrategie" bezeichnet wird, kann jedoch eine verkehrsunabhängige und fast optimale Speicherbelegungsgrenze m angegeben werden:

$$m = \frac{S}{\sqrt{N}} \quad , \tag{3.1}$$

wobei S die Gesamtanzahl der Speicherplätze, und N die Anzahl der Richtungen ist.

Auch bei dieser Organisationsform sind Verkehrsströme mit einem sehr niedrigen Verkehrswert im Überlastfall wesentlich benachteiligt. Dies kann verhindert werden durch eine Speicherplatzreservierung für jede Richtung (SMA). Eine verkehrsunabhängige heuristische Approximation für den optimalen Wert der Speicherplatzreservierung k pro Richtung lautet [Latouche (1980)]:

$$k = \frac{S}{N + \sqrt{N}} \tag{3.2}$$

Bei symmetrischen Verkehrswerten weisen die beiden Strategien SMXQ und SMA vergleichbare Resultate auf. Dagegen ist bei unsymmetrischen Verkehrsverhältnissen SMA wegen ihrer fairen Zuteilung der Betriebsmittel vorzuziehen. Die Kombinationen der beiden letzten Organisationsformen (SMQMA) gewährleistet jedoch den Schutz gegen Überlastsituationen. Bei dieser kombinierten Version können die Dimensionierungsparameter k und m nicht mit einer einfachen Regel angegeben werden.

Die Speicherorganisationsformen wurden für verschiedene Verkehrsverhältnisse und bezüglich verschiedener Leistungskriterien ausführlich untersucht in [Kamoun (1976), Kermani/Kleinrock (1977), Irland (1978), Kamoun/Kleinrock (1980), Latouche (1980), Körner (1983)].

b) Unterscheidung der Pakete nach bereits durchlaufenen Übertragungsabschnitten

Ein wichtiges Ziel einer Überlastabwehrstrategie besteht darin, die Entstehung von Blindlast durch Paketwiederholungen zu verhindern. Deshalb soll die Wahrscheinlichkeit für die Abweisung von Paketen mit zunehmendem Fortschreiten durch das Netz abnehmen. Dazu wird jedes Paket durch eine sogenannte Pufferklasse gekennzeichnet. Beim Betreten des Paketvermittlungsnetzes erhält es die Klasse 0 und nach jedem Übertragungsabschnitt wird die Klasse um eins erhöht. Mit jeder Erhöhung der Pufferklasse nimmt auch die Grenze für die maximale Speicherbelegung zu, so daß mit dieser Maßnahme der erzielte Effekt erreicht wird. Darüberhinaus wird auf diese Weise ein verklemmungsfreier Betrieb garantiert.

Dieses Pufferklassen-Konzept ist implementiert im GMD-Netz (Gesellschaft für Mathematik und Datenverarbeitung). Simulative Untersuchungen sind u.a. beschrieben in [Giessler/Hänle/König/Pade (1978), Giessler/Jägemann/Mäser/Hänle (1981)].

c) Unterscheidung der Pakete nach Ursprungs- oder Transitpaketen

Eine wirksame Methode, die das Entstehen einer Blindlast verhindern kann, ist die Abweisung von Paketen direkt am Netzzugang. Dies ist somit ein Grenzfall des vorangehenden Kriteriums

und wird im GMD-Netz mit Hilfe der Pufferklasse O verwirklicht.

Als Entscheidungskriterium für das Akzeptieren oder Abweisen von Paketen aus dem Anschlußnetz verwendet man:

- die Anzahl der Ursprungspakete im Netzknoten,
- die Gesamtanzahl der Pakete im Netzknoten.

Verschiedene analytische und simulative Untersuchungen haben die Wirksamkeit dieser Methode bestätigt [Price (1977), Lam/Reiser (1979), Schwartz/Saad (1979), Saad/Schwartz (1980), Kamoun (1981), Lam/Lien (1981)].

Diese Überlastabwehrstrategien werden in Kapitel 6 näher betrachtet. Insbesondere werden die dynamischen Eigenschaften untersucht.

d) Unterscheidung der Pakete nach Dialog- oder Stapelbetrieb

In Paketvermittlungsnetzen mit einer anwendungsorientierten Prioritätseinteilung kann das Abweisen von Paketen, die sich bereits im Netz befinden, vorwiegend vermieden werden, wenn eine Möglichkeit zur Auslagerung gegeben ist. Die kurzzeitige Auslagerung bezieht sich auf Stapelpakete, denn für sie kann eine größere Durchlaufzeit in Kauf genommen werden. Auf diese Weise wird bei Bedarf für einen kurzen Augenblick Speicherplatz für Pakete mit harten Zeitbedingungen zur Verfügung gestellt. Die Lastspitze kann somit vom Netzknoten selbst aufgefangen werden und wird nicht in unkontrollierter Weise auf Nachbarknoten ausgedehnt. Diese Überlastabwehrstrategie wird in Kapitel 7 vorgestellt und untersucht.

3.4.3.2 Datenflußsteuerung

Eine zweite wichtige Gruppe von Überlastabwehrstrategien beruht auf einer Datenflußsteuerung. Sie gehören somit zu den globalen Überlastabwehrstrategien. Mechanismen dieser Art verwenden entweder die Begrenzung der Paketzahl im betreffenden Netzteil (Fenstermechanismus, Transportberechtigung (Permit)) oder den Austausch von Meldungen (Start/Stop, Drosselung, Betriebsmittelreservierung). Je nach Umfang des gesteuerten oder geregelten

Bereiches bzw. Art der Methode kann eine weitere Unterteilung vorgenommen werden.

a) Datenflußsteuerung in virtuellen Verbindungen

Bezieht sich die Datenflußsteuerung auf eine einzige Verbindung (Transport Ebene), so vermag sie zwar die virtuelle Verbindung gegen eine exzessive Belastung zu schützen, kann aber wegen der vielen gleichzeitigen virtuellen Verbindungen das Paketvermittlungsnetz nicht vor Überlastsituationen bewahren. Dennoch übernimmt die individuelle Datenflußsteuerung eine wichtige Aufgabe, denn wenn bei mehreren virtuellen Verbindungen die Paketrate des Sendeprozesses größer ist als die des Empfangsprozesses, so ist das Paketvermittlungsnetz durch Rückstau schnell überlastet. Die Datenflußsteuerung begrenzt die Paketzahl pro virtuelle Verbindung, und ein momentan überlasteter Empfangsprozeß kann mit einer entsprechenden Meldung seinen Datenfluß vorübergehend stoppen. Darüberhinaus haben Quittierungsverzögerungen als Folge einer Netzüberlastung eine bremsende Wirkung auf den Datenfluß in den betreffenden virtuellen Verbindungen.

Auf diese Weise kann auch die individuelle Datenflußsteuerung dazu beitragen, daß die Überlastsituation sich nicht weiter zuspitzt. Andererseits können längere Quittierungsverzögerungen aber auch eine zusätzliche Netzbelastung durch Paketwiederholungen verursachen. Wenn jedoch der sendeseitige Netzendknoten über Informationen bezüglich Netzverzögerungen verfügt, können die individuellen Zeitüberwachungen auf der Transportebene des Anwenderprozesses durch entsprechende Meldungen adaptiv eingestellt werden.

Außer zwischen beiden Anwenderprozessen (DEE-DEE) kann die individuelle Datenflußsteuerung auch zwischen beiden Netzendknoten (DÜE-DÜE) vorkommen. Beispiele dafür sind GMD-Netz [Giessler/Hänle/König/Pade (1978)] und Datapac [Sproule/Mellor (1981)].

Bezüglich der analytischen bzw. simulativen Untersuchungen kann die individuelle Datenflußsteuerung als Spezialfall der Datenflußsteuerung zwischen zwei Netzendknoten betrachtet

werden. Die Datenflußsteuerung an der X.25 Netzschnittstelle wird in [Dieterle (1983)] durch Simulation untersucht.

b) Datenflußsteuerung zwischen zwei Netzendknoten

In diesem Falle wird der Gesamtfluß zwischen zwei Netzendknoten gesteuert. Dabei ist es unwesentlich, ob die einzelnen Pakete bereits zu einer individuellen Datenflußsteuerung gehören oder nicht. Die Betrachtungen gelten somit auch für den Datagrammbetrieb. Das Hauptziel besteht darin, Eintritts- und Austrittsrate für diesen Verkehrsstrom aufeinander abzustimmen. Bei einer Verstopfung der Verkehrsabgänge im entfernten Netzendknoten soll der Paketfluß im Ursprungsknoten gestoppt werden.

Die Mehrzahl dieser Datenflußsteuerungen basieren auf dem Fenstermechanismus. Maßgebende Parameter hierfür sind die Fenstergröße und das Time-Out-Intervall [Sunshine (1977), Pujolle (1979a), Kermani/Kleinrock (1980), Kleinrock/Kermani (1980), van As (1983)]. Die Datenflußsteuerung mit einem Fenstermechanismus ist in vielen Varianten analytisch untersucht worden: eine einzige Verkehrsbeziehung [Pennotti/Schwartz (1975), Schwartz/Saad (1979)], eine einzige Verkehrsbeziehung mit zufälliger Verkehrslenkung im Netz [Chatterjee/Georganas/Verma (1977)], mehrere Verkehrsbeziehungen [Reiser (1979)], eine Kombination von einer Ende-zu-Ende Datenflußsteuerung und einer globalen Datenflußsteuerung [Wong/Unsoy (1977)], eine Datenflußsteuerung auf drei Ebenen: global, Ende-zu-Ende und lokal [Georganas (1980)]. Das Problem der Einstellung der Zeitbegrenzung (Time-Out) wird in Kapitel 9 behandelt.

Außer dem Fenstermechanismus sind auch andere Verfahren möglich. In [Kleinrock/Tseng (1980)] wird eine Datenflußsteuerung untersucht, die auf einer kontrollierten Generierung von Transportberechtigungen (Permits) basiert. Das IBM spezifische Verfahren Pacing wird in [Schwartz (1982)] analysiert. Die Verwendung von Drosselungspaketen (choke-packets) wird in [Majitia/Irland/Grangé/Cohen/O'Donnell(1979)] an Hand eines Netzmodells simulativ untersucht. In [Matsumoto/Mori (1981)] erfolgt die Datenflußsteuerung durch Drosselung der Ankunftsrate. Der Einfluß einer stochastischen Verzögerung für eine derartige Datenflußsteuerung wird in Kapitel 8 betrachtet.

Vergleiche verschiedener Datenflußsteuerungen findet man in [Pujolle (1979b), Labetoulle/Pujolle (1981), Schwartz (1982)]. In [Irland/Pujolle (1980)] werden zwei Verfahren für die Paketwiederholung einander gegenübergestellt.

Mit Hilfe eines Netzsimulationsmodells wird in [Giessler/Jägemann/Mäser (1981)] die Ende-zu-Ende Datenflußsteuerung im Zusammenhang mit einem garantierten Durchsatz untersucht. In [Reiser (1981c)] steht insbesondere die Verzögerung, die Pakete im Anschlußnetz erfahren bis sie von der Datenflußsteuerung akzeptiert werden, im Vordergrund.

c) Datenflußsteuerung zwischen zwei benachbarten Netzknoten

Auch zwischen benachbarten Netzknoten ist eine Datenflußsteuerung erforderlich. Sie hat die Aufgabe, Überlastsituationen in einem Netzknoten durch Beschränkung der Menge der eintreffenden Pakete aus einem benachbarten Knoten Grenzen zu setzen. Je mehr Nachbarknoten existieren, um so schwieriger kann aber auf diese Weise eine Knotenüberlastung verhindert werden.

Als Realisierungsformen kommen der Fenstermechanismus, die Vergabe von Transportberechtigungen (Permits) oder der Austausch von Meldungen aufgrund von Schwellenwertüberschreitungen in Betracht. Ihre Realisierung mit einem Fenstermechanismus wird in [Pennotti/Schwartz (1975), Schwartz/Saad (1979), Georganas (1980)] analysiert. In [Kermani (1981)] wird die Ankunftsrate schrittweise geändert. [Harrison (1982)] verwendet eine Kombination verschiedener Verfahren (Fenster, Permits, Schwellenwert). In [Chu/Fayolle/Hibbits (1981), Kaufman/Gopinath/Wunderlich (1981)] wird insbesondere die enge Kopplung zwischen zwei benachbarten Netzknoten betrachtet.

d) Regelung der Gesamtbelastung des Netzes

Diese als isarithmetrische Datenflußsteuerung bezeichnete Methode - denn sie hält die Anzahl von Paketen im Netz konstant -, verwendet eine feste Anzahl von Transportberechtigungen (Permits), die durch die Paketübermittlung im Netz zirkulieren. Ein Paket kann nur dann befördert werden, wenn der Netzknoten über ein Permit verfügt. Das Paket führt dieses

Permit bis zum Zielknoten mit, wo es nach Abgabe dem Zielknoten zur Verfügung steht. Wegen Problemen wie die Häufung von Permits, geringer Durchsatz bei Stapelbetrieb, Möglichkeit zur lokalen Überlastung und Garantie, daß keine Permits verloren gehen sollen, hat diese Methode im wesentlichen nur noch eine historische Bedeutung.

Die Untersuchungen werden in [Davies (1972), Price (1977)] beschrieben. In [Takahashi/Shigeta/Hasegawa (1981)] wird eine Kombination von der isarithmetrischen Datenflußsteuerung und einer Regelung für den Netzzugang, die auf einer Speicherverwaltungsstrategie beruht, betrachtet. Die Kombination mit einer Datenflußsteuerung zwischen Netzendknoten wird in [Wong/Unsoy (1977)] untersucht. Eine Integration einer isarithmischen, einer Ende-zu-Ende sowie einer lokalen Datenflußsteuerung betrachtet [Georganas (1980)].

3.4.3.3 Ablaufsteuerung

Mit Hilfe der Ablaufsteuerung in den Netzknoten kann die Vermittlungsreihenfolge von Paketen, die zu verschiedenen Verkehrsströmen gehören, abgeändert werden. Dies ist der Fall bei Benutzerprioritäten (Dialog- oder Stapelbetrieb) und bei dynamischen Prioritäten (z.B. zuerst die Pakete, die am Ziel angekommen sind, dann die abgehenden Pakete und danach erst die neuen Pakete). Darüberhinaus können Pakete einzelner Verkehrsströme je nach Stau in den Ausgangswarteschlangen benachbarter Netzknoten zurückgestellt werden, so daß die Übertragungskapazität von Verkehrsströmen ohne nachfolgenden Stau genutzt werden kann. Eine derartige Betriebsweise setzt aber einen entsprechenden Meldungsaustausch zwischen den Netzknoten voraus. In [Majus (1981)] wird die Vermittlungsreihenfolge bestimmt durch einen Paketübermittlungstermin.

3.4.3.4 Verkehrslenkung

Die Verkehrslenkung dient dazu, den Paketverkehr möglichst gleichmäßig über das Netz zu verteilen, so daß eine optimale Betriebsmittelausnutzung angestrebt werden kann. Darüberhinaus erhöht sie bei Ausfällen von Übertragungsstrecken oder Netz-

knoten die Verfügbarkeit des Netzes durch Bereitstellung von Alternativwegen. Im Bezug auf Überlastabwehr ist die Verkehrslenkung als eine wichtige, aber langsam wirkende Maßnahme zu betrachten. In [McQuillan (1979)] werden die gegenseitigen Beziehungen zwischen Überlastabwehr und Verkehrslenkung diskutiert. Es wird ferner gezeigt, daß Verkehrslenkungsalgorithmen neben Messungen bezüglich Netzkapazität und Durchlaufzeit auch die implementierten Überlastabwehrmaßnahmen und insbesondere die Methode der Paketwiederholung berücksichtigen müssen. Die Frage wie dynamisch die Verkehrslenkung zu gestalten ist, wird in [Rudin/Mueller (1980)] mit Hilfe eines Simulationsmodells ausführlich untersucht. Die Resultate zeigen, daß eine adaptive Verkehrslenkung pro Paket zwar bei mittlerer Netzlast die Verkehrseigenschaften des Paketvermittlungsnetzes verbessert, jedoch bei höherer Belastung des Netzes eine Leistungsminderung bewirkt. Im Überlastfall bietet somit eine Verkehrslenkung ohne eine momentane Anpassung an die wechselnden Lastverhältnisse Vorteile. Dagegen weisen Lenkungsalgorithmen, die nur beim Verbindungsaufbau adaptiv sind, auch für höhere Belastungen gute Verkehrseigenschaften auf.

Falls aber das Verkehrsangebot stark unsymmetrisch oder nicht vorhersagbar ist, wird in [Chou/Bragg/Nilsson (1981)] die Notwendigkeit einer adaptiven Verkehrslenkung nachgewiesen. In [Yum (1981)] wird eine Netzkonfiguration mit mehreren Wegen zwischen Ursprung und Ziel betrachtet. Dabei wird gezeigt, daß bei Anwendung einer sequentiellen Wahl der verfügbaren Wege kürzere Durchlaufzeiten erreicht werden als im Falle einer zufälligen Auswahl. In [Yum/Schwartz (1981)] erfolgt zusätzlich noch eine lokale, adaptive Verkehrslenkungsentscheidung, die aufgrund der jeweiligen Warteschlangenlänge getroffen wird. Eine ähnliche Strategie wird in [Boorstyn/Livne (1981)] untersucht. In [Chu/Shen (1980)] wird eine nach dem Überlaufprinzip konzipierte hierarchische Verkehrslenkung untersucht. Als Entscheidungskriterium zum Wechsel vom Erstweg zum Alternativweg wird die Auslastung des Übertragungsabschnittes verwendet.

3.4.3.5 Topologie

Mit der Festlegung der geographischen Struktur des Paketvermittlungsnetzes werden gleichzeitig wesentliche Randbedingungen für die Überlastabwehr bestimmt. Dazu gehören vor allem die Netzvermaschung und die Netzhierarchie. Diese beiden Faktoren sind maßgebend für die Anzahl der direkten Wege, die Anzahl der alternativen Wege, die Anzahl der Übertragungsstrecken zwischen Ursprungs- und Zielknoten sowie die Netzrekonfigurationsmöglichkeiten bei Ausfällen von Betriebsmitteln. Die Netztopologie muß deshalb stets bei der Implementierung von Überlastabwehrmechanismen einbezogen werden.

3.4.3.6 Tarifgestaltung

Auch die Tarifgestaltung kann als Mittel zum Abbau von Betriebsmittelengpässen herangezogen werden. Falls Engpässe aufgrund von statistischen Auswertungen nach Art, Zeit und Ort festgestellt sind, läßt sich durch eine geeignete Gebührenpolitik eine zeitliche Verteilung des Verkehrsangebotes erreichen. Zum Beispiel durch Verlegung des Stapelbetriebes in die billigeren Nachtstunden.

3.5 Überlastproblematik in der Fernsprechvermittlung

Abschließend wird noch auf Literaturstellen ähnlicher Überlastprobleme in der Fernsprechvermittlung hingewiesen. Die Netzführung wird beschrieben in [Gimpelson (1974), Haenschke/Kettler/Oberer (1981)]. In [Lemieux (1981)] werden die analogen Eigenschaften der Netzführung in Fernsprech- und Paketvermittlungsnetzen diskutiert. In [Tran-Gia (1982)] wird die Überlastproblematik in rechnergesteuerten Fernsprechvermittlungssystemen sowie ihre Modellbildung und Analyse ausführlich behandelt.

4. MODELLIERUNG UND ANALYSEMETHODEN

Die Modellbildung spielt bei der Untersuchung komplexer Vorgänge in Paketvermittlungssystemen eine entscheidende Rolle [Kühn (1981)]. Je nach Betrachtung sind dabei zwei unterschiedliche, aber sich ergänzende Beschreibungsmodelle möglich:

- funktionsbezogene Modelle,
- verkehrsbezogene oder verkehrstheoretische Modelle.

Während bei den funktionsbezogenen Modellen das Hauptinteresse darin liegt, die einzelnen funktionellen Zusammenhänge zwischen Eingangs-, Zustands- und Ausgangsgrößen eines Systems zu beschreiben, interessiert bei den verkehrstheoretischen Modellen die Systemreaktion auf die Gesamtheit aller zufällig eintreffenden Anforderungen.

In Zusammenhang mit Überlastabwehr ist die funktionelle Beschreibungsweise äußerst wichtig für die Überprüfung von Protokollen auf Widerspruchs- und Verklemmungsfreiheit [Merlin (1979), Merlin/Schweitzer (1980), Günther (1981)].

Darüberhinaus setzt eine wirklichkeitsnahe verkehrstheoretische Modellbildung gute Kenntnisse der funktionellen Abläufe voraus.

Im weiteren werden verkehrstheoretische Modelle betrachtet, die sich dazu eignen, Überlastsituationen in Paketvermittlungsnetzen quantitativ zu erfassen und die Wirksamkeit der eingesetzten Überlastabwehrstrategien zu untersuchen.

In diesem Kapitel werden deshalb zuerst die verschiedenen Komponenten eines verkehrstheoretischen Modells vorgestellt und die wichtigsten charakteristischen Verkehrsgrößen definiert. Für die Modelluntersuchung stehen analytische und simulative Methoden zur Verfügung. Die analytischen Methoden, die in dieser Arbeit verwendet werden, basieren auf der Theorie der Markoff-Prozesse. Einerseits handelt es sich hierbei um die Auflösung von Differentialgleichungssystemen bzw. linearen Gleichungssystemen, die aufgrund eines sogenannten Markoff-Zustandsdiagramms aufgestellt werden, andererseits aber werden die Eigenschaften von Warteschlangennetzen mit Produktlösungs-

form ausgenutzt. Weitere analytische Verfahren werden in der umfangreichen Literatur über Warteschlangentheorie (Bedienungstheorie, Verkehrstheorie, queueing) behandelt, z.B. [Takács (1962), Cohen (1969/1982), Cooper (1972/1981), Gross/Harris (1974), Kleinrock (1975, 1976), Allen (1978)]. Die bei analytischen Methoden oft zwingenden Modellvereinfachungen (Unabhängigkeitsannahmen, spezielle Verteilungsfunktionen, Strukturen oder Betriebsorganisationen) können durch simulative Methoden umgangen werden. Somit ist auch die Verkehrssimulation ein bedeutendes Hilfsmittel zur Untersuchung von Verkehrsmodellen.

Für die Auswahl der Methode sind nachfolgende Kriterien zu berücksichtigen:

Analytische Methoden:

- Lösungsweg nicht immer von vornherein bekannt,
- Modellvereinfachungen notwendig,
- exakter oder approximativer Lösungsweg möglich,
- einfache Parameterstudien,
- vorwiegend kleine Rechenzeiten.

Simulative Methoden:

- fast beliebiger Detaillierungsgrad,
- durch die vielen Einzelheiten und Systemzusammenhänge oft schwierige Interpretation der Resultate,
- Resultate als statistische Ergebnisse,
- Programmierung meistens aufwendig,
- aufwendige Parameterstudien,
- große Rechenzeiten.

4.1 Verkehrstheoretische Modellbildung

Um die Verkehrsmodelle einfach und eindeutig behandeln zu können, müssen die komplexen Vorgänge und Zusammenhänge in den realen Systemen abstrahiert werden.

Für ein Paketvermittlungsnetz kommen die folgenden Gesichtspunkte als Modellkomponenten in Betracht:

- Netztopologie
- Struktur der Netzknoten
- Prozessoren (Vermittlung, Leitungsmodulen)
- Speicher
- Leitungen
- Kommunikationsprotokolle (Datenflußsteuerung, Überlastabwehr)
- Betriebsorganisation (Prioritäten, Zuteilungs- und Abfertigungsstrategien)
- Charakteristik der Verkehrsquellen (Dialogbetrieb, Stapelbetrieb)
- Verkehrsbeziehungen.

Je nach Ziel und Durchführung einer Untersuchung müssen die relevanten Aspekte in einem verkehrstheoretischen Modell berücksichtigt werden.

4.1.1 Ein allgemeines einstufiges Warteschlangensystem

Als Basis für komplexere Modelle wird nach Bild 4.1 ein allgemeines einstufiges Warteschlangensystem als Grundmodell betrachtet. Dieses Modell repräsentiert zum Beispiel einen Leitungsmodul in einem modularen Multiprozessor-Netzknoten oder einen Netzknoten selbst, in dem der Zeitanteil der inneren Organisation gegenüber der Übertragungsgeschwindigkeit der Leitungen vernachlässigt werden kann. Da es sich bei diesem Modell nicht um eine spezielle Anwendung handeln soll, werden in der nachfolgenden Beschreibung die allgemeinen Begriffe aus der Warteschlangentheorie verwendet: Anforderung, Bedienungseinheit, usw..

Ohne vorerst auf Einzelheiten einzugehen, kann das Ablaufgeschehen in diesem Verkehrsmodell wie folgt beschrieben werden. Eintreffende Anforderungen kommen entweder aus Verkehrsquellen oder von anderen Warteschlangensystemen. Je nach Systemzustand und Systemzugangskriterien können diese Anforderungen abgewie-

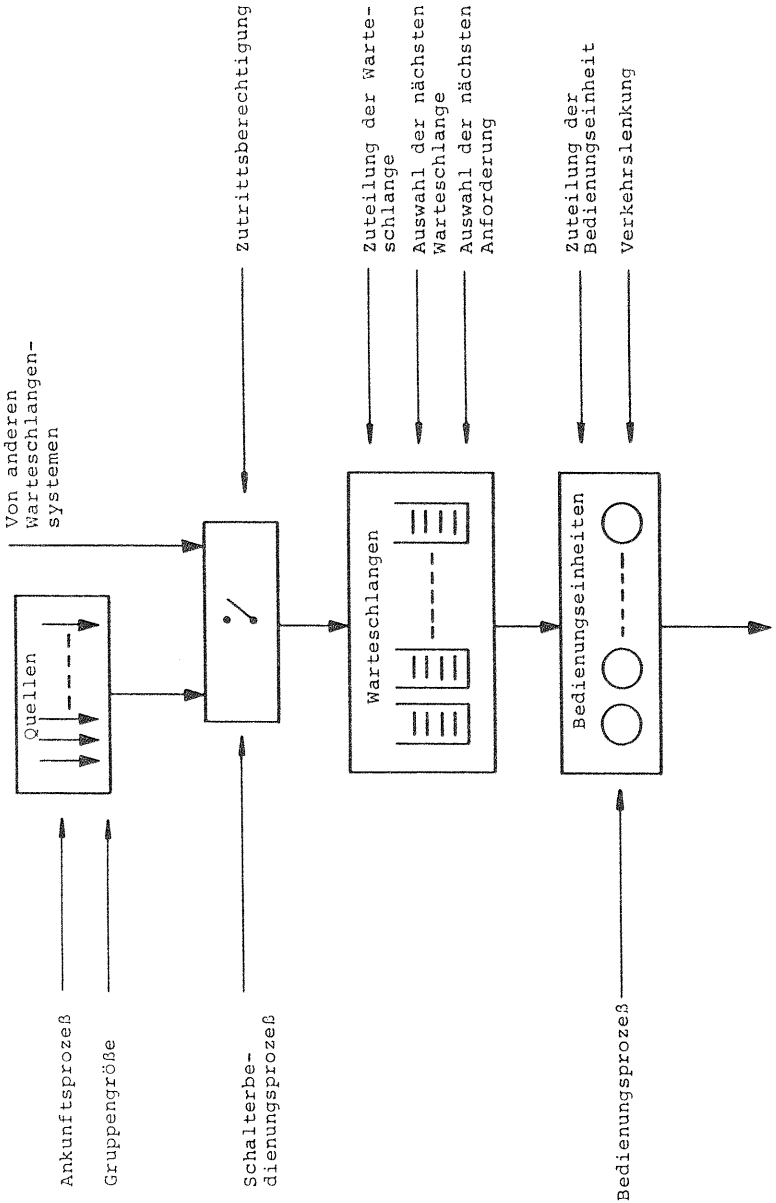


Bild 4.1: Allgemeines einstufiges Warteschlangensystem.

sen werden. Dies ist durch einen Schalter symbolisiert. Sind bei Eintreffen einer Anforderung noch Bedienungseinheiten frei, so wird eine davon belegt. Andernfalls wird die Anforderung in eine der Warteschlangen eingereiht. Bei jedem Bedienungsende wird die gerade bediente Anforderung weitergeleitet und verläßt somit das System. Entsprechend der Abfertigungsstrategie wird eine der wartenden Anforderungen zur Bedienung ausgesucht.

Zur vollständigen Charakterisierung von Verkehrsmodellen ist eine Reihe von Parametern hinsichtlich Struktur, Betriebsweise und Verkehr notwendig.

4.1.2 Strukturparameter

Die Struktur eines modellierten Paketvermittlungsnetzes oder Netzknotens besteht im allgemeinen aus einem Netz von einstufigen Warteschlangensystemen. Die Netztopologie bestimmt die Verbindungswege zwischen diesen sogenannten Grundmodellen, die selbst gekennzeichnet sind durch die Zahl und Anordnung ihrer Strukturkomponenten.

Als Strukturparameter sind hier zu nennen:

- Verkehrsquellen (sources), in denen Anforderungen in zufälligen Abständen erzeugt werden,
- Verkehrseingänge (ports), zur Aufnahme von Anforderungen aus anderen Grundmodellen,
- Schalter (switches) zur Unterbrechung des Anforderungsstromes,
- Warteschlangen (queues) zur Zwischenspeicherung von eintreffenden Anforderungen,
- Bedienungseinheiten (servers) für die Ausführung der Bedienungsanforderungen.

4.1.3 Betriebsparameter

Als Betriebsparameter gelten alle Angaben, die die Betriebsweise kennzeichnen. Dazu gehören:

- Zuteilungsstrategien zur Auswahl einer Bedienungseinheit der gleichen Kategorie (service unit allocation), z.B. zufällige, sequentielle oder zyklische Zuteilung.
- Zuteilungsstrategien zur Auswahl der Warteschlange (queue allocation), z.B. nach Priorität der eintreffenden Anforderung, nach gewünschter Richtung, nach kürzester Warteschlangenlänge, nach einer zufälligen, sequentiellen oder zyklischen Zuteilung.
- Abfertigungsstrategien zur Auswahl der nächsten abzufertigenden Warteschlange (interqueue discipline), z.B. nach Prioritäten, zufällig, sequentiell oder zyklisch.
- Abfertigungsstrategien zur Auswahl der nächsten Anforderung innerhalb einer Warteschlange (queue discipline), z.B. nach der Reihenfolge des Eintreffens (FIFO, First In, First Out), zufällig (RANDOM) oder in inverser Reihenfolge des Eintreffens (LIFO, Last In, First Out).
- Prioritätsstrategien zur Behandlung von Anforderungen unterschiedlicher Dringlichkeit (priority discipline), z.B. unterbrechende-, nichtunterbrechende Prioritäten oder deren Kombination.
- Behandlung eintreffender Anforderungen in Blockierungsfällen (access strategies), z.B. Abweisung mit oder ohne spätere Wiederholmöglichkeit bei vollem Speicher oder Abweisung bei Überschreitung einer vorgegebenen Grenze für die betreffende Kategorie von Anforderungen.
- Verkehrslenkung von Anforderungen (routing), z.B. feste, zufällige, alternative oder adaptive Verkehrslenkung.

4.1.4 Verkehrsparameter

Das Verkehrsgeschehen selbst wird beschrieben durch die geographischen Verkehrsbeziehungen (Verkehrsmatrix) und durch die statistischen Eigenschaften der Ankunftszeitpunkte von Anforderungen und deren Bedienungszeiten. In der Regel haben die Ankunftsabstände und die Bedienungszeiten einen zufälligen Charakter, und somit entspricht der zeitliche Ablauf innerhalb des betrachteten Systemmodells einem stochastischen Prozeß (Zufallsprozeß).

Als Ausgangspunkt dienen folgende Prozesse:

- Ankunftsprozeß (arrival process), beschrieben durch die Verteilung der Ankunftsabstände der Anforderungen als unabhängige und identisch verteilte Zufallsvariable T_A (interarrival time). Die Verteilungsfunktion

$$A(t) = P\{T_A \leq t\} \tag{4.1}$$

ist definiert als die Wahrscheinlichkeit, daß der Ankunftsabstand T_A höchstens gleich der beliebigen Zeit t ist.

Bei einem mittleren Ankunftsabstand t_A wird der Kehrwert als mittlere Ankunftsrate definiert:

$$\lambda = \frac{1}{t_A} \quad . \tag{4.2}$$

- Bedienungsprozeß (service process), beschrieben durch die Verteilung der Bedienungszeiten der einzelnen Anforderungen als unabhängige und identisch verteilte Zufallsvariable T_H (holding time, service time). Mit μ als mittlere Bedienungsrate und h als mittlere Bedienungsdauer gilt entsprechend

$$H(t) = P\{T_H \leq t\} \quad \text{und} \quad \mu = \frac{1}{h} \quad . \tag{4.3}$$

- Gruppenankunftsprozeß oder Gruppenbedienungsprozeß. Bei dieser Art von Prozessen muß neben den Ankunftsabständen bzw. den Bedienungsdauern die Verteilung der Gruppengröße (batch size) zusätzlich noch angegeben werden.

Diese Verkehrsprozesse bestimmen zusammen mit den Struktur- und Betriebsparametern den Systemablauf. Somit können daraus folgende Zufallsprozesse abgeleitet werden:

- Systemzustandsprozeß für die zufallsabhängige Anzahl X von Anforderungen im System,
- Warteprozeß für die zufallsabhängigen Wartezeiten T_W (Waiting time),
- Durchlaufprozeß für die zufallsabhängigen Durchlaufzeiten T_F (Flow time),

- Ausgangsprozeß für die zufallsabhängigen Ausgangsabstände T_D (Departure time).

Je nach Wahl der Verteilungsfunktion für die Ankunfts- und Bedienungsprozesse unterscheidet man zwischen Markoff-Prozessen und Nicht-Markoff-Prozessen.

Bei Markoff-Prozessen hängt der zukünftige Prozeßverlauf nur vom momentanen Zustand des Prozesses ab. Diese Gedächtnislosigkeit wird als Markoff-Eigenschaft bezeichnet. Nicht-Markoff-Prozesse, bei denen der weitere Prozeßverlauf zusätzlich noch von der Vergangenheit abhängt, können dann auf einen Prozeß mit der Markoff-Eigenschaft zurückgeführt werden, wenn im Ablauf des Prozesses Zeitpunkte mit der Eigenschaft der Gedächtnisfreiheit definiert werden können (eingebettete Markoff-Kette, Phasenmethode, Methode der Gedächtnisvariablen).

Zur Charakterisierung von Warteschlangenmodellen wird generell eine erweiterte Form der Kendallschen Klassifikation GI/G/n/S verwendet:

GI : Typ des Ankunftsprozesses
(GI=General Independent, allgemein verteilter Zwischenankunftsabstand)
G : Typ des Bedienungsprozesses
(G = General, allgemein verteilte Bedienungszeit)
n : Anzahl der Bedienungseinheiten
S : Kapazität des Warteschlangensystems
(Warteplätze und Bedienungseinheiten)

Die wichtigsten Verteilungsfunktionen für GI bzw. G, die in verkehrstheoretischen Modellen Anwendung finden, sind:

M : negativ exponentielle Verteilungsfunktion (Markoff-Prozeß),
D : konstante Verteilungsfunktion (Deterministischer Prozeß),
E_k : Erlang-Verteilungsfunktion k-ter Ordnung,
H_k : Hyperexponentielle Verteilungsfunktion k-ter Ordnung.

4.1.5 Charakteristische Verkehrsgrößen

Ziel der verkehrstheoretischen Untersuchungen ist es, die charakteristischen Verkehrsgrößen zu bestimmen, um auf diese Weise die Zusammenhänge zwischen Struktur, Betriebsweise und Verkehrsangebot erfassen zu können.

Wichtige Kriterien für die Güte der Verkehrsabwicklung in Paketvermittlungsnetzen sind:

- Verlustwahrscheinlichkeit B (loss probability), definiert als die Wahrscheinlichkeit, daß ein eintreffendes Paket abgewiesen wird und dadurch entweder verloren geht oder später wiederholt werden muß.
- Blockierungswahrscheinlichkeit P_B (blocking probability), definiert als die Wahrscheinlichkeit, daß ein bestimmtes Betriebsmittel blockiert ist, also nicht belegt werden kann.
- mittlere Systembelastung E[X] (mean system occupancy), definiert als die mittlere Anzahl von Paketen, die sich gleichzeitig im betrachteten System befinden.
- mittlere Wartebelastung Ω (mean queue length), definiert als die mittlere Anzahl von Paketen, die vor einer Bedienungseinheit auf Bearbeitung warten (aktive Wartebelastung).
- mittlere Speicherbelastung (mean buffer occupancy), definiert als die mittlere Anzahl von Paketen, die sich im Speicher befinden. Diese Belastung setzt sich zusammen aus denjenigen Paketen, die zu quittieren sind (passive Wartebelastung) und denjenigen, die vor der Bedienungseinheit warten (aktive Wartebelastung).
- mittlere Durchlaufzeit t_F (mean flow time), definiert als Summe aller Warte-, Bedienungs- und Transportzeiten.
- mittlerer Durchsatz D (throughput), definiert als die Anzahl von Paketen, die im Mittel pro Zeiteinheit vom betrachteten System bearbeitet werden.
- Leistung P (power), definiert als Verhältnis von mittlerem Durchsatz zu mittlerer Durchlaufzeit.

- Systemerholungszeit (system recovery time), definiert als die mittlere Zeit, die ein System nach einer Überlastsituation benötigt, um seinen stationären Zustand - bis auf eine spezifizierte Abweichung - wieder zu erreichen.
- Paketwiederholungsrate (packet repetition rate), definiert als die Frequenz, mit der Pakete wiederholt werden müssen.

In vielen Fällen der Praxis geben die charakteristischen Mittelwerte und Wahrscheinlichkeiten genügend Auskunft über das Verkehrsverhalten eines Systems. Mitunter sind aber diese Angaben nicht ausreichend. In diesem Falle sind die Verteilungsfunktionen zu bestimmen. Ein typisches Beispiel ist die Durchlaufzeitverteilungsfunktion, die auch in dieser Arbeit betrachtet wird.

4.2 Beschreibung des Ablaufgeschehens von verkehrstheoretischen Modellen mit einem Markoff-Prozeß

Grundlage der wahrscheinlichkeitstheoretischen Methoden zur Untersuchung von verkehrstheoretischen Modellen sind die stochastischen Prozesse, unter denen die diskreten Markoff-Prozesse mit einem kontinuierlichen Zeitparameter eine Sonderstellung einnehmen [Cox/Miller (1965), Feller (1966, 1968), Çinlar (1975), Mehdi (1981), Heyman/Sobel (1982), Ross (1983)].

Die Sonderstellung dieses Markoff-Prozesses ist auf die folgenden Merkmale zurückzuführen:

- Im Gegensatz zu allgemeineren Prozessen, die sich in der Regel nur für einfach strukturierte Modelle verwenden lassen, können die gedächtnisfreien Markoff-Prozesse auch kompliziertere Modelle, die für Untersuchungen bezüglich Überlastabwehrstrategien erforderlich sind, adäquat beschreiben.
- Für verschiedene Überlastabwehrstrategien ist ihr dynamisches Verhalten entscheidend, so daß eine instationäre Analysemethode notwendig ist. Mit einem Markoff-Prozeß kann ein instationärer Vorgang mathematisch behandelt werden.

- Falls erforderlich lassen sich auch allgemeinere Prozesse mit der Phasenmethode (Erlang-k-Verteilung, hyperexponentielle Verteilung, Cox-Verteilung [Kleinrock (1975)] oder empirische Verteilung [Bux/Herzog (1977a, 1977b), Bux (1979)]) gut annähern. Dies ist auch in instationärem Fall möglich [Odoni/Roth (1983)].
- Eine Prozeßverallgemeinerung mit Hilfe einer eingebetteten Markoff-Kette kann sowohl stationär [Kleinrock (1975)] als auch instationär erfolgen [Tran-Gia (1982)].
- Die mehrdimensionale Beschreibung eines Modells mit einem Markoff-Prozeß läßt sich weitgehend automatisieren und numerisch gut auswerten (vgl. Abschnitte 4.4 und 4.6).

In diesem Abschnitt werden die mathematischen Grundlagen zur Anwendung der diskreten Markoff-Prozesse mit einem kontinuierlichen Zeitparameter t in kurzer Form dargestellt. Insbesondere wird gezeigt, daß nicht nur der Systemzustandsprozeß, sondern auch der Warteprozeß bzw. der Durchlaufprozeß instationär behandelt werden kann.

4.2.1 Der Markoff-Prozeß

Ein zeitkontinuierlicher stochastischer Prozeß mit diskreten Werten $\{X(t) = x,\ t \geq 0,\ x \in M\}$, M = Zustandsraum, ist ein Markoff-Prozeß, wenn der weitere Verlauf dieses Prozesses nur vom gegenwärtigen Zustand x_n zum Betrachtungszeitpunkt t_n abhängt und nicht von der Vorgeschichte.

Für den wahrscheinlichen Zustand zum nächstbetrachteten Zeitpunkt t_{n+1} gilt somit:

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

$$\text{mit} \quad t_0 < t_1 < \ldots < t_n < t_{n+1} \quad \text{und} \quad \{x_1, x_2, \ldots, x_{n+1} \in M\} \tag{4.4}$$

Auf Grund der Gedächtnislosigkeit ist das stochastische Verhalten eines Markoff-Prozesses zu einem beliebigen Zeitpunkt t eindeutig bestimmt durch eine Anfangsverteilung der Prozeßzustände und die Wahrscheinlichkeiten für die Zustandsübergänge. Es kann

gezeigt werden, daß die negativ exponentielle Verteilung als einzige kontinuierliche Verteilungsfunktion die sogenannte Markoff-Eigenschaft besitzt, und die zeitlichen Abstände zwischen zwei Zustandsänderungen sind deshalb auch stets negativ exponentiell verteilt. Im allgemeinen ist der betrachtete Prozeß durch einen mehrdimensionalen Zustand charakterisiert. Der Prozeß wird dann mit Hilfe eines Prozeßzustandsvektors $\underline{X}(t)$ beschrieben.

4.2.2 Die Kolmogoroffschen Differentialgleichungen

Ein Markoff-Prozeß $\{X(t),\ t \geq 0\}$ sei zum Zeitpunkt s im Zustand i und zu einem späteren Zeitpunkt t im Zustand j. Die Wahrscheinlichkeit für diesen Zustandsübergang ist definiert durch die Übergangswahrscheinlichkeit

$$p_{ij}(s,t) = P\{X(t) = j \mid X(s) = i\} \quad , \quad s < t \quad . \tag{4.5}$$

Durch Betrachtung eines Zustandes $X(u) = k$ zu einem beliebigen Zwischenzeitpunkt u gibt die Chapman-Kolmogoroff-Gleichung einen Zusammenhang zwischen der Übergangswahrscheinlichkeit des gesamten Zustandsübergangs von i nach j und den Übergangswahrscheinlichkeiten der Zwischenübergänge $p_{ik}(s,u)$ und $p_{kj}(u,t)$.

$$p_{ij}(s,t) = \sum_k p_{ik}(s,u) \cdot p_{kj}(u,t) \quad , \quad s < u < t \quad . \tag{4.6}$$

Wählt man für den Zwischenzeitpunkt $u=t-\Delta t$, so bekommt man für Gl.(4.6):

$$\begin{aligned} p_{ij}(s,t) &= \sum_k p_{ik}(s,t-\Delta t) \cdot p_{kj}(t-\Delta t,t) \\ &= p_{ij}(s,t-\Delta t) \cdot p_{jj}(t-\Delta t,t) + \sum_{k \neq j} p_{ik}(s,t-\Delta t) \cdot p_{kj}(t-\Delta t,t) \quad . \end{aligned} \tag{4.7}$$

Durch Subtraktion mit dem Term $p_{ij}(s,t-\Delta t)$ auf beiden Seiten und anschließender Division durch Δt ergibt sich:

$$\frac{p_{ij}(s,t)-p_{ij}(s,t-\Delta t)}{\Delta t} = -\frac{1-p_{jj}(t-\Delta t,t)}{\Delta t} \cdot p_{ij}(s,t-\Delta t) + \sum_{k \neq j} p_{ik}(s,t-\Delta t) \cdot \frac{p_{kj}(t-\Delta t,t)}{\Delta t} \quad . \tag{4.8}$$

Bildet man den Grenzwert $\Delta t \to 0$, erhält man die Kolmogoroff-Vorwärts-Differentialgleichung für die Übergangswahrscheinlichkeiten:

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

wobei gilt:

$$q_j(t) = \lim_{\Delta t \to 0} \frac{1-p_{jj}(t-\Delta t,t)}{\Delta t} \quad \text{und} \quad q_{kj}(t) = \lim_{\Delta t \to 0} \frac{p_{kj}(t-\Delta t,t)}{\Delta t} \quad ; \quad k \neq j \quad . \tag{4.9}$$

Wird andererseits der Zwischenpunkt $u=s+\Delta s$ in Gl.(4.6) eingesetzt, erhält man:

$$\begin{aligned} p_{ij}(s,t) &= \sum_k p_{ik}(s,s+\Delta s) \cdot p_{kj}(s+\Delta s,t) \\ &= p_{ii}(s,s+\Delta s) \cdot p_{ij}(s+\Delta s,t) + \sum_{k \neq i} p_{ik}(s,s+\Delta s) \cdot p_{kj}(s+\Delta s,t) \quad ; \quad s < u < t \ . \end{aligned} \tag{4.10}$$

Durch Subtraktion mit dem Term $p_{ij}(s+\Delta s,t)$ auf beiden Seiten und Division durch Δs bekommt man:

$$\frac{p_{ij}(s,t)-p_{ij}(s+\Delta s,t)}{\Delta s} = -\frac{1-p_{ii}(s,s+\Delta s)}{\Delta s} \cdot p_{ij}(s+\Delta s,t) + \sum_{k \neq i} \frac{p_{ik}(s,s+\Delta s)}{\Delta s} \cdot p_{kj}(s+\Delta s,t) \tag{4.11}$$

Durch Bildung des Grenzwertes $\Delta s \to 0$ ergibt sich die Kolmogoroff-Rückwärts-Differentialgleichung für die Übergangswahrscheinlichkeiten:

$$\frac{d}{ds} p_{ij}(s,t) = -q_i(s) \cdot p_{ij}(s,t) + \sum_{k \neq i} q_{ik}(s) \cdot p_{kj}(s,t) \quad , \quad s < t$$

mit

$$q_i(s) = \lim_{\Delta s \to 0} \frac{1-p_{ii}(s,s+\Delta s)}{\Delta s} \quad \text{und} \quad q_{ik}(s) = \lim_{\Delta s \to 0} \frac{p_{ik}(s,s+\Delta s)}{\Delta s} \quad , \quad k \neq i \quad . \tag{4.12}$$

$q_j(t)$, $q_{kj}(t)$ bzw. $q_i(s)$, $q_{ik}(s)$ werden als Übergangswahrscheinlichkeitsdichten oder Übergangsraten bezeichnet.

4.2.3 Systemzustandsprozeß

Grundlegend für die Analyse eines Verkehrsmodells sind die Zustandswahrscheinlichkeiten des Zustandsraumes M zu einer beliebigen Zeit t

$$P_j(t) = P\{X(t) = j\} \quad . \tag{4.13}$$

Diese Zustandswahrscheinlichkeiten lassen sich aus einer Anfangsverteilung zur Zeit s=0 und den Übergangswahrscheinlichkeiten nach dem Gesetz der totalen Wahrscheinlichkeit berechnen:

$$P_j(t) = \sum_i P_i(0) \cdot P_{ij}(0,t) \quad . \tag{4.14}$$

Ferner gilt die Normierungsbedingung

$$\sum_j P_j(t) = 1 \quad . \tag{4.15}$$

Wird (4.9) in der Gleichung (4.14) eingesetzt, erhält man die Kolmogoroff-Vorwärts-Differentialgleichung für die Berechnung der Zustandswahrscheinlichkeiten zum Zeitpunkt t

$$\frac{d}{dt} P_j(t) = -q_j(t) \cdot P_j(t) + \sum_{k \neq j} q_{kj}(t) \cdot P_k(t) \quad . \tag{4.16}$$

Sind im Spezialfall die Übergangsraten zeitunabhängig, so spricht man von einem homogenen Markoff-Prozeß.

Ist außerdem das System im eingeschwungenen Zustand

$$P_j = P_j(\infty) = \lim_{t \to \infty} P_j(t) \tag{4.17}$$

erhält man aus der Kolmogoroff-Vorwärts-Differentialgleichung (4.16) mit

$$\lim_{t \to \infty} \frac{d}{dt} P_j(t) = 0 \tag{4.18}$$

ein System von linearen Gleichungen

$$q_j P_j = \sum_{k \neq j} q_{kj} \cdot P_k \quad . \tag{4.19}$$

In diesem statistischen Gleichgewichtszustand ist also die Wahrscheinlichkeitsrate für das Verlassen des Zustandes j gleich der Wahrscheinlichkeitsrate des Entstehens dieses Zustandes aus allen benachbarten Zuständen.

Zum Aufstellen der Differentialgleichungen für die Zustandswahrscheinlichkeiten ist das Zustandsdiagramm, aus dem sämtliche Zustandsübergänge und deren Übergangsraten hervorgehen, ein sehr anschauliches Hilfsmittel.

Wird der Zustand j betrachtet, so wird die zugehörige Differentialgleichung gemäß Gl.(4.16) aufgestellt, wobei die folgenden Regeln gelten:

- die Übergangsrate $q_j(t)$ ist die Summe aller Übergangsraten, die den betrachteten Zustand j verlassen (Summe der Raten sämtlicher abgehender Pfeile),
- jede Übergangsrate $q_{kj}(t)$, $k \neq j$, ist die Übergangsrate vom Nachbarzustand k zum betrachteten Zustand j (jeweils der betreffende ankommende Pfeil).

Die Auflösung dieses Differentialgleichungssystems ergibt unter Berücksichtigung der Anfangsbedingungen $P_j(0)$ die zeitabhängigen Zustandswahrscheinlichkeiten $P_j(t)$, $j \in M$.

Bild 4.2: Zustandsdiagramm für das Warteschlangenmodell M/M/1/S.

Beispiel

Als einfaches Beispiel sei das Warteschlangensystem M/M/1/S betrachtet. Das Zustandsdiagramm zeigt Bild 4.2. Der Systemzustand $X(t) = j$, $j = 0,1,\ldots,S$ bezeichne die Anzahl von Anforderungen im System zum Zeitpunkt t. Bei zeitabhängigen Übergangsraten (Ankunftsrate $\lambda(t)$, Bedienungsrate $\mu(t)$) entwickelt sich ein Systemzustandsprozeß gemäß den folgenden Kolmogoroff-Vorwärts-Differentialgleichungen:

$$\frac{d}{dt} P_0(t) = -\lambda(t) \cdot P_0(t) + \mu(t) \cdot P_1(t)$$

$$\frac{d}{dt} P_j(t) = -[\lambda(t) + \mu(t)] \cdot P_j(t) + \lambda(t) \cdot P_{j-1}(t) + \mu(t) \cdot P_{j+1}(t) \quad , \quad 1 < j < S$$

$$\frac{d}{dt} P_S(t) = -\mu(t) \cdot P_S(t) + \lambda(t) \cdot P_{S-1}(t) \quad . \tag{4.20}$$

4.2.4 Der Warte- und Durchlaufprozeß

Zur Berechnung von Wartezeit- oder Durchlaufzeitverteilungsfunktionen in komplexen Markoff-Systemen oder in Markoff-Systemen mit zeitabhängigen Übergangsraten kommt dem Konzept des Warte- oder Durchlaufprozesses eine maßgebene Bedeutung zu [Syski (1964), Kühn (1972, 1983)]. Der Unterschied zwischen einem Warteprozeß und einem Durchlaufprozeß besteht darin, daß in letzterem Falle die Bedienung selbst auch zum Prozeß gehört. Im weiteren wird der Durchlaufprozeß betrachtet.

Die Bestimmung der Durchlaufzeit erfolgt mit einer sogenannten Testanforderung. Diese Testanforderung trifft zu einem beliebigen Zeitpunkt s ein und, falls sie nicht vom System abgewiesen wird, beginnt ihr Durchlaufprozeß. Auf diese Weise kann unter Berücksichtigung jeder Beeinflussung das Schicksal dieser Testanforderung genau verfolgt werden. Der Durchlaufprozeß endet, wenn ein sogenannter absorbierender Zustand erreicht wird. Im Gegensatz zum Systemzustandsprozeß hat der Durchlaufprozeß also eine endliche Dauer.

Für den Durchlaufprozeß bezüglich einer eintreffenden Testanforderung gelten folgende, vom Zustandsprozeß im allgemeinen abweichende Bezeichnungen bzw. Definitionen:

$\hat{M}$: Menge der sogenannten Durchlaufzustände der Testanforderung. Die Zustände können unterteilt werden in Zustände der Mengen $\hat{M}_1$ und $\hat{M}_2$.

$\hat{M}_1$: Menge sämtlicher Systemzustände in denen der Durchlaufprozeß beginnen kann. Diese Zustände sind für eine eintreffende Testanforderung direkt erreichbar und sie sind eine Untermenge des Zustandsraumes M.

$\hat{M}_2$: Menge zusätzlich definierter Durchlaufzustände, die während des Durchlaufprozesses eingenommen werden können. Diese Zustände sind für eine eintreffende Testanforderung nur indirekt erreichbar.

H : Menge der absorbierenden Zustände für den Durchlaufprozeß (Bedienungsende oder Verlust).

$\hat{p}_{ij}(s,t)$: Übergangswahrscheinlichkeit des Durchlaufprozesses von einem beliebigen Beginnzustand i (direkt oder indirekt erreichbar), der von der Testanforderung zum Zeitpunkt s eingenommen wird, nach Zustand j zum Zeitpunkt t; $i,j \in \hat{M}$.

$\hat{q}_i(s)$, $\hat{q}_{ij}(s)$: Übergangsraten bezogen auf den Durchlaufprozeß; $i,j \in \hat{M}$.

$T_F(s)$: zufällige Durchlaufzeit einer Testanforderung, die zum Zeitpunkt s ihren Durchlaufprozeß beginnt.

$P_i(s)$: Antreffwahrscheinlichkeit für den Systemzustand i zum Zeitpunkt s; $i \in \hat{M}_1$ (nur direkt erreichbare Zustände).

Um den Durchlaufprozeß formal zu beschreiben, wird die bedingte komplementäre Durchlaufzeitverteilungsfunktion $f_i(s,t)$ eingeführt. Sie ist definiert als die Wahrscheinlichkeit, daß die zufällige Durchlaufzeit $T_F(s)$ länger als die Zeit $t^* = t-s$ beträgt, falls die Testanforderung ihren Durchlaufprozeß zum Zeitpunkt s im Zustand i beginnt:

$$f_i(s,t) = P\{T_F(s) > t^* \mid \text{Beginnzustand } i\} \quad , \quad i \in \hat{M} \quad , \quad t = s+t^* \quad . \tag{4.21}$$

Diese Durchlaufzeit ist länger als t^*, wenn die Testanforderungen während des Intervalls $[s,s+t^*]$ außerhalb der absorbierenden Menge H bleibt. Deshalb kann man die bedingte komplementäre Durchlaufzeitverteilungsfunktion als Summe von Übergangswahrscheinlichkeiten ausdrücken:

$$f_i(s,t) = \sum_{j \notin H} \hat{p}_{ij}(s,t) \quad . \tag{4.22}$$

Durch Summation von Gl.(4.12) gemäß Gl.(4.22) erhält man die Kolmogoroff-Rückwärts-Differentialgleichung für die bedingte komplementäre Durchlaufzeitverteilungsfunktion $f_i(s,t)$:

$$\frac{d}{ds} f_i(s,t) = -\hat{q}_i(s) \cdot f_i(s,t) + \sum_{\substack{k \neq i \\ k \notin H}} \hat{q}_{ik}(s) \cdot f_k(s,t) \quad . \tag{4.23}$$

Als Anfangsbedingung gilt, daß jede Testanforderung, die nicht abgewiesen wird (also erfolgreich ist), eine positive Durchlaufzeit mit Wahrscheinlichkeit 1 erfährt:

$$f_i(s,s) = 1. \tag{4.24}$$

Die Auflösung dieses Systems von Differentialgleichungen ergibt unter Berücksichtigung der Anfangsbedingungen $f_i(s,s) = 1$ die bedingten komplementären Durchlaufzeitverteilungsfunktionen $f_i(s,t)$, $i \in \hat{M}$.

Die Beschreibung des Durchlaufprozesses durch die Kolmogoroff-Rückwärts-Differentialgleichung bezieht sich somit auf einen Beginnzustand und diejenigen Nachbarzustände, die durch einen einzigen Übergang aus dem Beginnzustand erreicht werden können. Der Durchlaufprozeß wird mit Hilfe von einem Durchlaufzustandsdiagramm dargestellt. Dieses Diagramm enthält sowohl sämtliche Zustände, die eine Testanforderung während ihres Durchlaufprozesses einnehmen kann, als auch deren Übergänge mit den zugehörigen Übergangsraten.

Die Durchlaufzustandsdiagramme in dieser Arbeit zeigen die Zustandsmuster, die für den Beginnzustand des Durchlaufprozesses in Frage kommen. Diese Zustandsmuster enthalten die Testanforderung selbst nicht. Es muß so aufgebaut sein, daß sämtliche anderen Anforderungen, die die Durchlaufzeit der betrachteten Anforderung beeinflussen können, berücksichtigt werden. Das Beenden des Durchlaufprozesses ist jeweils durch einen dick gezeichneten Übergangspfeil zur absorbierenden Zustandsmenge H gekennzeichnet. Testanforderungen, die ein volles System antreffen, werden abgewiesen und gehen sofort verloren. Die Zustände für ein volles System sind schraffiert gezeichnet.

Aufgrund des Durchlaufzustandsdiagrammes werden die Differentialgleichungen für die bedingten komplementären Durchlaufzeitverteilungsfunktionen $f_i(s,t)$ jeweils gemäß Gl.(4.23) aufgestellt. Dabei gelten für den Zustand i folgende Regeln:

- nur die Übergänge, die den betrachteten Zustand i verlassen, müssen berücksichtigt werden (nur abgehende Pfeile),
- die Übergangsrate $\hat{q}_i(s)$ ist die Summe aller Übergangsraten, die den betrachteten Zustand i verlassen (Summe der Raten sämtlicher abgehender Pfeile),
- jede Übergangsrate $\hat{q}_{ik}(s)$, $k \neq i$, ist die Übergangsrate vom betrachteten Zustand i zum direkt erreichbaren Nachbarzustand k (jeweils der betreffende abgehende Pfeil).

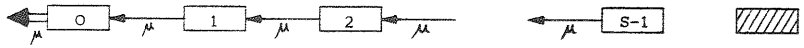


Bild 4.3: Durchlaufzustandsdiagramm für das Warteschlangenmodell M/M/1/S.

Beispiel

Bild 4.3 zeigt das Durchlaufzustandsdiagramm für das System M/M/1/S. Der Durchlaufprozeß der Testanforderung kann außer bei Antreffen eines vollen Systems (Zustand S) in jedem Systemzustand begonnen werden und endet nach der Bedienung der Testanforderung, dargestellt durch Zustand 0.

Mathematisch gilt das nachfolgende System von Kolmogoroff-Rückwärts-Differentialgleichungen.

$$\frac{d}{ds} f_0(s,t) = -\mu(s) \cdot f_0(s,t) \tag{4.25}$$

$$\frac{d}{ds} f_i(s,t) = -\mu(s) \cdot f_i(s,t) + \mu(s) \cdot f_{i-1}(s,t) \quad , \quad i = 1,\ldots,S-1$$

Die komplementäre Durchlaufzeitverteilungsfunktion selbst - für eine zum Zeitpunkt s eintreffende Testanforderung - läßt sich durch Gewichtung mit den Antreffwahrscheinlichkeiten $P_i(s)$ ermitteln:

$$F^C(s,t) = \sum_i P_i(s) \cdot f_i(s,t) \quad , \quad i \in \hat{M}_1 \quad . \tag{4.26}$$

Aus der komplementären Durchlaufzeitverteilungsfunktion wird die mittlere Durchlaufzeit $t_F(s)$ einer zum Zeitpunkt s eintreffenden Testanforderung durch Integration bestimmt:

$$t_F(s) = \int_{\tau=s}^{\infty} F^C(s,\tau)\,d\tau \quad . \tag{4.27}$$

4.2.5 Charakteristische Verkehrsgrößen

Ausgehend von den stationären oder transienten Zustandswahrscheinlichkeiten können die wichtigsten charakteristischen Verkehrsgrößen berechnet werden. Die Herleitungen für die einzelnen Größen sind modellabhängig. Sie werden vorteilhaft anhand des betreffenden Systemzustandsraumes durchgeführt. In vielen Markoff-Modellen wird das analytische Ergebnis einer charakteristischen Verkehrsgröße ähnlich hergeleitet. Als erstes Beispiel sind hier einige wichtige Größen für das Warteschlangenmodell M/M/1/S zusammengestellt. Charakteristische Verkehrsgrößen für komplexere Beispiele sind jeweils im Zusammenhang mit den untersuchten Modellen angegeben.

a) Verlustwahrscheinlichkeit B(t):

$$B(t) = P\{\text{Verlust}\} = P_S(t) \quad . \tag{4.28}$$

b) Wartewahrscheinlichkeit W(t):

$$\begin{aligned} W(t) &= P\{\text{Bedienungseinheit belegt und Anforderung wird nicht abgewiesen}\} \\ &= 1-P_O(t)-P_S(t) \end{aligned} \tag{4.29}$$

c) Mittlere Systembelastung E[X(t)]:

Erwartungswert aller Anforderungen im System,

$$E[X(t)] = \sum_{j=O}^{S} j \cdot P_j(t) \tag{4.30}$$

d) Mittlere Wartebelastung $\Omega(t)$:

Erwartungswert der belegten Warteplätze,

$$\Omega(t) = \sum_{j=1}^{S} (j-1) \cdot P_j(t) \tag{4.31}$$

e) Momentane Abgangsrate oder Durchsatz D(t):

$$\begin{aligned} D(t) &= P\{\text{Bedienungseinheit belegt}\} \cdot \mu(t) \\ &= \sum_{j=1}^{S} P_j(t) \cdot \mu(t) = [1-P_O(t)] \cdot \mu(t) \end{aligned} \tag{4.32}$$

f) Mittlere Durchlaufzeit $t_F(s)$:

Bei einer zeitinvarianten mittleren Bedienungsdauer $h=1/\mu$ und FIFO-Abfertigungsstrategie läßt sich die mittlere Durchlaufzeit einer beliebigen, zum Zeitpunkt s eintreffenden Anforderung auf folgende Weise aus dem angetroffenen Systemzustand ermitteln:

- Werden j Anforderungen von der Testanforderung vorgefunden, so beträgt ihre mittlere Durchlaufzeit (j+1)h.
- Wird ein volles System vorgefunden, so wird die Testanforderung abgewiesen. Dieser Fall liefert somit keinen Beitrag zur Durchlaufzeit.
- Durch Gewichtung der bedingten Durchlaufzeiten mit den Antreffwahrscheinlichkeiten erhält man die mittlere Durchlaufzeit:

$$\begin{aligned} t_F(s) &= \sum_{j=O}^{S-1} P_j(s) \cdot (j+1)h + P_S(s) \cdot O \\ &= h \cdot \left[E[X(s)] - (S+1) \cdot P_S(s) + 1\right] \quad . \end{aligned} \tag{4.33}$$

Bei einer zeitabhängigen Bedienungsrate $\mu(t)$ oder bei anderen Abfertigungsstrategien muß die mittlere Durchlaufzeit $t_F(s)$ mit Hilfe eines Durchlaufprozesses bestimmt werden (Vgl. Abschnitt 4.2.4).

4.3 Warteschlangennetze mit Produktlösungsform

Die Modellierung der Datenflußsteuerung bzw. globaler Überlastabwehrstrategien in Paketvermittlungsnetzen führt oft auf ein Warteschlangennetz. Hierunter versteht man die topologische Verknüpfung einzelner Warteschlangensysteme. Die exakte Berechnung einer relativ allgemeinen Klasse von Warteschlangennetzen - die sogenannten BCMP-Netze [Baskett/Chandy/Muntz/Palacios (1975)] - basiert auf dem Produktform-Ansatz. Bedingt durch diesen Ansatz können viele Warteschlangensysteme in einem Verkehrsmodell berücksichtigt werden. Somit kann die sequentielle Beanspruchung von Bedienungseinheiten durch eine Anforderung sowie deren gleichzeitige Beanspruchung durch verschiedene Anforderungen in das Verkehrsablaufgeschehen einbezogen werden. Allerdings erfüllt nicht jedes Modell genau die Voraussetzungen für eine Produktlösungsform. In diesem Fall ist man auf Approximationsverfahren angewiesen: Blockierungen im Netz [Takahashi/Miyahara/Hasegawa (1980)], Netze mit allgemeinen Warteschlangensystemen [Kühn (1979), Whitt (1983a, 1983b)], Netze mit Prioritäten [Schmitt (1983)].

In den folgenden Abschnitten werden die Möglichkeiten der BCMP-Netze nur insoweit ausgeschöpft, als es für die Modellierung von Datenflußsteuerung und Überlastabwehr in Paketvermittlungsnetzen notwendig ist. Für eine ausführliche Darstellung über weitere Modellierungsaspekte wird verwiesen auf [Reiser (1982), Wong/Lam (1982), Lam/Wong (1982)]. Die allgemeine Theorie der Warteschlangennetze wird zum Beispiel in [Gelenbe/Mitrani (1980), Sauer/Chandy (1981), Lavenberg (1983)] beschrieben.

4.3.1 Voraussetzungen

Die Modellierung von Paketvermittlungsnetzen als Warteschlangennetz mit einer Produktformlösung basiert im wesentlichen auf folgenden Voraussetzungen:

- Die einzelnen virtuellen Verbindungen bzw. die logischen Verknüpfungen zwischen Netzendknoten entsprechen verschiedenen Klassen von Anforderungen, wobei alle Klassen, die den gleichen Weg durchlaufen, zu einer Kette zusammengefaßt werden. Sind c_k Klassen in der Kette k vorhanden und existieren K Ketten im Netz, so beträgt die Anzahl von Klassen $c = c_1+c_2+\ldots+c_K$.
- Das Paketvermittlungsnetz kann als offenes, geschlossenes oder gemischtes Warteschlangennetz modelliert werden. Externe Ankünfte erfolgen bei offenen Netzen entsprechend exponentiell verteilter Ankunftsabstände. Die Gesamtankunftsrate $\lambda_0 = \lambda_0(x)$ darf dabei abhängig von der Gesamtzahl x von Anforderungen im Netz sein. Bei geschlossenen Netzen zirkulieren in jeder Klasse eine konstante Anzahl von Anforderungen.
- Die Übertragungsabschnitte werden als FIFO-Warteschlangensysteme mit nur einer Bedienungseinheit dargestellt. Die Übertragungsgeschwindigkeit für den Übertragungsabschnitt i beträgt C_i bit/s. Sie muß für alle Klassen identisch sein.
- Die wesentlich schnelleren Prozessoren in den Netzknoten werden im allgemeinen als Verzweigungspunkte modelliert. Anderenfalls verkörpern sie FIFO-Warteschlangensysteme mit jeweils einer einzigen Bedienungseinheit. In diesem Fall ist die klassenunabhängige Arbeitsgeschwindigkeit des Prozessors i gleich C_i bit/s.
- Stochastische Verzögerungen für Quittungen und Time-Outs werden mit Hilfe von einem Warteschlangensystem mit unendlich vielen Bedienungseinheiten (IS-System, Infinite-Server-System) nachgebildet. Die Verzögerung des IS-Systems i wird charakterisiert durch eine Verzögerungsrate C_{ik} bit/s, die ketten-(klassen-)abhängig sein darf.
- Die Verkehrslenkung, die deterministisch oder stochastisch sein darf, ist durch eine Verzweigungsmatrix für jede Kette k festgelegt:

$$\underline{Q}^{(k)} = \left(q_{ij}^{(k)}\right) \qquad (4.34)$$

dabei ist $q_{ij}^{(k)}$ die Verzweigungswahrscheinlichkeit einer Anforderung der Kette k von System i nach System j.

- Die Paketlänge $1/\mu$ ist negativ exponentiell verteilt, wobei diese Verteilung für alle Klassen identisch ist. Bei einer Übertragungsgeschwindigkeit bzw. einer Verzögerungsrate von C bit/s gehorchen auch die Bedienungszeiten einer negativ exponentiellen Verteilung. Die mittlere Bedienungszeit beträgt:

$$h = \frac{1}{\mu C} \tag{4.35}$$

- Die Länge eines Paketes wird während seines Netzdurchlaufes nicht beibehalten, sondern jedes Paket hat aufgrund der Unabhängigkeitsannahme von Kleinrock [Kleinrock (1964)] in jedem Warteschlangensystem eine andere Länge. Nur durch diese Annahme können Paketvermittlungsnetze analytisch behandelt werden.

4.3.2 Zustandswahrscheinlichkeiten

Betrachtet wird ein Warteschlangennetz, das aus N Warteschlangensystemen und K Ketten besteht. Einfachheitshalber enthält jede Kette genau eine Klasse von Anforderungen.

Es wird die folgende Notation für die Kette k verwendet:

λ_{0k} : externe Ankunftsrate

λ_{ik} : Ankunftsrate in System i

$q_{ij}^{(k)}$: Verzweigungswahrscheinlichkeit von System i nach System j

A_{ik} : Verkehrsangebot für System i

C_{ik} : Verzögerungsrate für IS-System i.

Ferner bedeutet:

$1/\mu$: mittlere Paketlänge

A_i : Gesamtverkehrsangebot für System i

C_i : Übertragungs- bzw. Arbeitsgeschwindigkeit für FIFO-System i.

Bezeichnet X_i die Gesamtanzahl von Anforderungen in System i, so kennzeichnet der Zustandsvektor $\underline{X} = (X_1, X_2, \ldots X_N)$ den Zustand des Netzes. Unter den erwähnten Voraussetzungen läßt sich die Wahrscheinlichkeit für das Auftreten dieses Zustandes im stationären Fall als Produkt von unabhängigen Zustandswahrscheinlichkeiten der einzelnen Warteschlangensysteme $i = 1,\ldots,N$, ausdrücken:

$$P(x_1, x_2, \ldots, x_N) = \frac{1}{G} \cdot \prod_{i=1}^{N} P_i(x_i) \quad . \tag{4.36}$$

wobei

$$P_i(x_i) = \begin{cases} A_i^{x_i} & \text{für FIFO-Systeme} \\ \dfrac{A_i^{x_i}}{x_i!} & \text{für IS-Systeme} \quad . \end{cases} \tag{4.37}$$

Die Normierungskonstante G berechnet sich für offene Warteschlangennetze als Produkt von Einzeltermen, die unabhängig vom Zustand $\underline{X}$ sind:

$$G = \prod_{i=1}^{N} G_i$$

wobei (4.38)

$$G_i = \begin{cases} \dfrac{1}{1-A_i} & \text{für FIFO-Systeme} \\ e^{A_i} & \text{für IS-Systeme} \quad . \end{cases}$$

Bei geschlossenen Warteschlangennetzen ist die Normierungskonstante G abhängig von der Gesamtanzahl von Anforderungen x im Netz. Sie berechnet sich durch Summation von sämtlichen Kombinationen von Zustandswahrscheinlichkeiten $P(x_1, x_2, \ldots, x_N)$ des Zustandsraumes M mit $x = x_1 + x_2 + \ldots + x_N$ Anforderungen:

$$G = G(x) = \sum_{\underline{x} \in M} P(x_1, x_2, \ldots, x_N) \quad , \quad \underline{x} = (x_1, x_2, \ldots, x_N) \tag{4.39}$$

Bei gemischten Warteschlangennetzen wird die Normierungskonstante G bestimmt durch ein Produkt aus einer Normierungskonstanten G^O für alle offenen Ketten und einer Normierungskonstanten $G^C(x)$ für alle geschlossenen Ketten:

$$G = G^O \cdot G^C(x) \quad . \tag{4.40}$$

G^O berechnet man gemäß Gl.(4.38), wobei A_i ersetzt wird durch

$$A_i^O = \sum_{k \in o} A_{ik} \quad , \quad o: \{\text{alle offenen Ketten}\} \quad . \tag{4.41}$$

$G^C(x)$ erhält man gemäß Gl.(4.39), wobei A_i in Gl.(4.37) ersetzt wird durch:

$$A_i^C = \begin{cases} \sum\limits_{k \in c} \dfrac{A_{ik}}{1-A_i^O} & \text{für FIFO-Systeme} \\ \sum\limits_{k \in c} A_{ik} & \text{für IS-Systeme,} \end{cases} \tag{4.42}$$

$c: \{\text{alle geschlossenen Ketten}\}$.

Für die Berechnung der Verkehrsangebote A_i werden zuerst die Ankunftsraten λ_{ik} bestimmt. Dies geschieht mit Hilfe des Verkehrsflußgleichgewichts für die Kette k, wobei die externe Rate λ_{Ok} für eine geschlossene Kette gleich Null ist. Somit entsteht das folgende Gleichungssystem für die Kette k:

$$\lambda_{ik} = \lambda_{Ok} \cdot q_{Oi}^{(k)} + \sum_{j=1}^{N} \lambda_{jk} \cdot q_{ji}^{(k)} \quad , \quad i = 1,2,\ldots,N \quad . \tag{4.43}$$

Daraus ergibt sich für die kettenabhängigen Verkehrsangebote A_{ik}:

$$A_{ik} = \begin{cases} \dfrac{\lambda_{ik}}{\mu C_i} & \text{für FIFO-Systeme} \\ \dfrac{\lambda_{ik}}{\mu C_{ik}} & \text{für IS-Systeme} \quad . \end{cases} \tag{4.44}$$

Die gesuchten Verkehrsangebote A_i werden dann durch entsprechende Summation erhalten:

$$A_i = \sum_{k=1}^{K} A_{ik} \tag{4.45}$$

Bei gemischten Netzen werden die offenen Ketten bzw. die geschlossenen Ketten getrennt aufsummiert.

4.3.3 Charakteristische Verkehrsgrößen

Aus den stationären Zustandswahrscheinlichkeiten lassen sich die interessierenden charakteristischen Verkehrsgrößen berechnen. Es handelt sich hierbei hauptsächlich um die Verlustwahrscheinlichkeit B bzw. die mittlere Systembelastung E[X], die beide durch die entsprechende Summation der Zustandswahrscheinlichkeiten bestimmt werden, sowie die daraus abgeleitete Verkehrsgröße Durchsatz D bzw. Durchlaufzeit t_F.

Der Durchsatz bei verlustbehafteten Systemen wird berechnet aus der Beziehung:

$$D = \lambda \cdot (1-B) \tag{4.46}$$

Die mittlere Durchlaufzeit errechnet sich mit dem Theorem von Little [Little (1961)]:

$$E[X] = \lambda \cdot t_F \tag{4.47}$$

Die charakteristischen Verkehrsgrößen beziehen sich auf das ganze Netz, ein einziges Warteschlangensystem, eine Kette oder eine Klasse. Im allgemeinen kann die Summation direkt im Berechnungsalgorithmus für die Zustandswahrscheinlichkeiten bzw. für die Normierungskonstante durchgeführt werden. Für offene Netze können die Verkehrsgrößen oft ohne Berechnung der Zustandswahrscheinlichkeiten aus bekannten Ergebnissen des M/M/1-Systems [Kleinrock (1975)] berechnet werden.

Bei Modellen für Speicherorganisationen, die meistens auf offenen Warteschlangennetzen mit einer Begrenzung der Anzahl der Anforderungen in jeder Kette beruhen [Lam (1977)], wird eine

begrenzte , mehrdimensionale Zustandsverteilung durch einen geeigneten, modellabhängigen Algorithmus normiert. Die Summe der Zustandswahrscheinlichkeiten, die zur Grenzfläche der k-ten Dimension gehören, ergibt die Verlustwahrscheinlichkeit B_k für die entsprechende Kette k [Irland (1978), Kamoun/Kleinrock (1980), Latouche (1980), Kaufman (1981), Körner (1983)].

Für geschlossene Netze existieren heute einige leistungsfähige Algorithmen [Buzen (1973), Reiser (1977, 1981a), Chandy/Sauer (1980), Sauer/Chandy (1981), Lavenberg (1983), Willmann (1983)]. Aufgrund des hohen Speicherbedarfs und der beträchtlichen Rechenzeiten bei Modellen mit vielen geschlossenen Ketten, die bei Netzuntersuchungen mit vielen virtuellen Verbindungen auftreten, sind spezielle Verfahren entwickelt worden [Reiser (1979), Chandy/Neuse (1982), Lam/Lien (1983)].

4.4 Numerische Methoden zur Lösung von Markoff-Modellen

Die Methoden zur numerischen Auswertung von Markoff-Modellen können eingeteilt werden nach dem zeitlichen Verlauf des stochastischen Prozesses (stationär, transient), nach den Eigenschaften des Zustandsdiagrammes (Produktformlösung, beliebig) oder nach der Art der Durchführung (Auswertung einer geschlossenen Lösung oder einer Transformierten; direkte, iterative oder rekursive numerische Methode). Aus dieser Vielfalt werden die numerischen Methoden behandelt, die in späteren Kapiteln eine Anwendung finden.

4.4.1 Iterative numerische Methoden

Iterative numerische Methoden bilden ein leistungsfähiges und universelles Instrument zur Lösung der stationären Kolmogoroff-Zustandsgleichungen von komplexen mehrdimensionalen Markoff-Systemen. In Bezug auf die erreichbare Modellierungstiefe sind von der Methode her keine Grenzen gesetzt. Sie wird lediglich eingeschränkt durch Kriterien wie Speicherbedarf, Rechenzeit und nicht zuletzt durch den Aufwand zur Umsetzung des Modells in ein lauffähiges Rechenprogramm.

Zum Vorbereitungsaufwand gehört:

- die Bestimmung sämtlicher Zustände und deren Übergänge,
- die Bestimmung der Übergangsraten,
- das Aufstellen der verschiedenen Zustandsgleichungen,
- die Verifikation, daß jeder abgehende Übergang einem ankommenden Übergang entspricht oder daß keine Übergänge vergessen worden sind (Dies hat eine wesentliche Bedeutung für mehrdimensionale Zustandsräume, die nicht mehr zusammenhängend darzustellen sind).

Eine Software-Unterstützung zur Reduktion dieser langwierigen Prozedur kann hier Abhilfe schaffen. Erste Schritte in dieser Richtung sind bereits gemacht worden [Müller (1980)].

Im allgemeinen werden die Punktiterationsverfahren nach Gauß-Seidel oder die Überrelaxationsverfahren (SOR, Successive Overrelaxation) zur Lösung der Zustandsgleichungen verwendet [Cooper (1972/1981)]. Neuerdings aber werden auch die Blockiterationsverfahren (Block-Gauß-Seidel, Block-SOR) und spezielle Konvergenzkriterien (Tschebyscheff-Beschleunigung, Polynom-Beschleunigung) vorteilhaft für die Lösung von großen Zustandsräumen eingesetzt [Courtois (1977), Müller (1980), Kaufman/Gopinath/Wunderlich (1981)].

In dieser Arbeit wurde das SOR-Punktiterationsverfahren angewendet. Beginnend mit einer passenden Anfangsverteilung der Zustandswahrscheinlichkeiten $P_j^{(0)}$, $j = 0,\ldots,N$, werden diese Zustandswahrscheinlichkeiten mit Hilfe einer Iterationsvorschrift immer wieder verbessert. Für die Zustandswahrscheinlichkeit P_j im (m+1)-ten Iterationsschritt gilt:

$$P_j^{(m+1)} = (1-\omega)\cdot P_j^{(m)} - \frac{\omega}{q_j}\cdot\left[\sum_{k=0}^{j-1} q_{kj}\cdot P_k^{(m+1)} + \sum_{k=j+1}^{N} q_{kj}\cdot P_k^{(m)}\right] \quad , \; j = 0,\ldots,N \tag{4.48}$$

Wie aus Gl.(4.48) ersichtlich ist, werden in diesem Verfahren die im betrachteten Iterationsdurchgang bereits verbesserten Näherungswerte berücksichtigt.

ω ist der sogenannte Relaxationsfaktor, mit dem man die Iterationskonvergenz beschleunigen kann. Dieser Faktor muß für jedes Verkehrsmodell empirisch bestimmt werden. In der Regel gilt als Richtwert $1.1 < \omega < 1.4$. Für $\omega = 1$ erhält man den Gauß-Seidel Algorithmus.

Die Iteration wird mit Hilfe einer geeigneten Konvergenzbedingung abgebrochen. Ein oft verwendetes Kriterium hierfür ist

$$\sum_j |P_j^{(m+1)} - P_j^{(m)}| < \varepsilon \qquad (4.49)$$

mit $\varepsilon = 10^{-6}, 10^{-8}, \ldots$.

Als Alternative kann man zur Berechnung der stationären Zustandswahrscheinlichkeiten auch das entsprechende System von Differentialgleichungen verwenden. Ausgehend von einem leeren System wird dann der zeitliche Systemverlauf mit einem transienten Verfahren solange verfolgt, bis die geforderte Stationäritätsgenauigkeit (entspricht der Konvergenzbedingung bei Iterationsverfahren) erreicht worden ist. Im Falle, daß keine gute Anfangsverteilung für die Zustandswahrscheinlichkeiten bekannt ist, ist dieses Verfahren in den meisten Fällen schneller. Dies hängt damit zusammen, daß ein leeres System eine natürliche Anfangsverteilung für den transienten Systemverlauf darstellt.

4.4.2 Rekursive numerische Methoden

Wie in [Herzog/Woo/Chandy (1975)] gezeigt wurde, läßt sich eine umfangreiche Klasse von Warteschlangensystemen durch ein System von rekursiven Gleichungen beschreiben. Der Grundgedanke dabei ist, daß bei einer geeigneten Wahl von n bekannt angenommenen Zustandswahrscheinlichkeiten P_X^γ, $\gamma = 1,\ldots,n$, sämtliche, beispielsweise durchnumerierten Zustandswahrscheinlichkeiten P_j, $j = 0,\ldots,N$, als Linearkombination dieser sogenannten Rekursionsstartpunkte ausgedrückt werden können. Für die Zustandswahrscheinlichkeit P_j gilt somit

$$P_j = C_j^1 \cdot P_X^1 + C_j^2 \cdot P_X^2 + C_j^3 \cdot P_X^3 + \ldots + C_j^n \cdot P_X^n \quad . \qquad (4.50)$$

In dieser Linearkombination ist der Koeffizient C_j^γ genau der zu P_X^γ gehörende Anteil von P_j. Die Koeffizienten C_j^γ erhält man durch n-fache Durchrechnung der Rekursionsgleichungen, wobei jedesmal ein anderer Startpunkt aktiv ist. Insbesondere bekommt man die Koeffizienten C_j^1 mit dem Startvektor (V,0,0,0...), die Koeffizienten C_j^2 mit dem Startvektor (0,V,0,0,...), usw.. Dabei ist V ein willkürlicher Startwert, den man jedoch in einem für die Rekursion günstigen numerischen Bereich wählt, oft gilt $V = 1$.

Als nächstes wird ein Rekursionsstartpunkt, zum Beispiel P_X^1, ausgewählt und die verbleibenden (n-1) Startpunkte als Funktion dieser unabhängigen Variablen ausgedrückt:

$$P_X^\gamma = f(P_X^1) \quad , \quad \gamma = 2,\ldots,n \quad . \qquad (4.51)$$

Zur Durchführung dieses Schrittes wird ein System von (n-1) unabhängigen Gleichungen zwischen Zustandswahrscheinlichkeiten P_j aufgestellt. Wird für jedes P_j in diesen Gleichungen die Gl.(4.50) angesetzt, erhält man (n-1) voneinander unabhängige Gleichungen für die restlichen Startpunkte $P_X^\gamma, \gamma = 2,\ldots,n$. Dieses üblicherweise wesentlich kleinere Gleichungssystem kann durch Matrix-Inversion oder mit Hilfe des Gauß-Algorithmus gelöst werden [Rutishauser (1976a), Maron (1982)].

Als nächster Schritt liefert die Gl.(4.50) - bis auf einen Faktor P_X^1 - alle Zustandswahrscheinlichkeiten P_j, so daß nach anschließender Normierung der Zustandsverteilung auch dieser Faktor P_X^1 bestimmt werden kann:

$$P_X^1 = \frac{V}{\sum_j P_j} \qquad (4.52)$$

Aus den Zustandswahrscheinlichkeiten P_j können die charakteristischen Verkehrsgrößen errechnet werden. Im Falle, daß man lediglich an charakteristischen Verkehrsgrößen interessiert ist, kann die entsprechende Summation der Zustandswahrscheinlichkeiten laufend vorgenommen werden und die berechneten Zustandswahrscheinlichkeiten brauchen nicht zusätzlich abgespeichert werden. Zum Zwecke der Normierung der charakteristischen Verkehrsgrößen werden dann alle Zustandswahrscheinlichkeiten nur aufsummiert.

Dieses Verfahren wurde für die stationäre Lösung von den Modellen in Kapitel 8 verwendet und wird dort ausführlich dargestellt. Mit einer Modifikation dieses Originalalgorithmus kann in einigen Fällen eine Verringerung der Rechenzeit und des Speicherbedarfs erzielt werden. Ein entsprechendes Beispiel wird in Kapitel 7 für ein Warteschlangensystem mit 2 nichtunterbrechenden Prioritäten und begrenztem Warteraum behandelt.

In diesem modifizierten Algorithmus werden alle Zustandswahrscheinlichkeiten nach Möglichkeit sofort in Termen eines einzigen Rekursionsstartpunktes P_X^1 ausgedrückt. Müssen bei der Aufstellung des Rechenalgorithmus zusätzlich Startpunkte P_X^γ $\gamma = 2,3,..$ eingeführt werden, dann werden sie, sobald sich eine geeignete Relation finden läßt, wieder eliminiert. Der geringe Speicherplatzbedarf gegenüber dem Originalalgorithmus kommt im wesentlichen dadurch zustande, daß die Zustandswahrscheinlichkeiten höchstens einmal abgespeichert werden müssen anstatt n-fach. Die Rechenzeitersparnis hängt zusammen mit der geringeren Anzahl von Rekursionsoperationen.

Die hier dargelegte rekursive Methode ist ein universelles Verfahren, das erfolgreich zur Lösung einer großen Klasse von mehrdimensionalen Markoff-Warteschlangensystemen mit begrenztem Warteraum eingesetzt werden kann. Ein anderes Verfahren basiert auf der sogenannten regenerativen Methode, mit der eine Reihe von eleganten und insbesondere stabilen rekursiven Algorithmen zur Lösung von Warteschlangensystemen mit unendlichem Warteraum entstanden ist [Tijms/van Hoorn (1981), van Hoorn (1981, 1983)].

Allgemein sind rekursive Algorithmen geprägt durch ihre Eleganz, Rechenzeit- und Speichereffizienz. Die erreichbare Genauigkeit der Resultate hängt aber sehr stark vom Algorithmus und von den Modellparametern ab. Im Gegensatz zu den iterativen Methoden kann also die Genauigkeit nicht gesteuert werden. Die Genauigkeit der errechneten Zustandswahrscheinlichkeiten kann durch Einsetzen in die Zustandsgleichungen (4.19) überprüft werden. Für die Genauigkeitsfehler der Zustandswahrscheinlichkeiten P_j gilt:

$$\delta = \left| P_j - \frac{1}{q_j} \cdot \sum_{k \neq j} q_{kj} \cdot P_k \right| \qquad (4.53)$$

Falls die geforderte Genauigkeit nicht erreicht wird , kann sie mit Hilfe der iterativen Methode verbessert werden (Nachiteration).

4.4.3 Numerische Methoden für transiente Vorgänge

Die Basis für die Auswertung von transienten Vorgängen in Markoff-Modellen sind die Kolmogoroff-Differentialgleichungen, die mit numerischen Methoden zur Lösung von gewöhnlichen Differentialgleichungen behandelt werden können. Speziell wurde das Runge-Kutta-Verfahren 4. Ordnung verwendet [White/Schmidt/Bennett (1975), Rutishauser (1976b), Maron (1982)]. In Bezug auf Rechenzeit, Speicherbedarf und Genauigkeit eignet sich dieses Verfahren sehr gut für die Behandlung von großen Systemen von Differentialgleichungen, wie sie bei der Behandlung von Markoff-Modellen auftreten.

Mit diesem Verfahren lassen sich sowohl die transienten Zustandswahrscheinlichkeiten $P_j(t)$ nach Gl.(4.16) als auch die bedingten komplementären Durchlaufzeitverteilungsfunktionen $f_i(s,t)$ nach Gl.(4.23) berechnen.

Betrachtet man das gekoppelte System von Differentialgleichungen für die Zustandswahrscheinlichkeiten, so werden nach dem Runge-Kutta-Verfahren 4. Ordnung die Zustandswahrscheinlichkeiten $P_j(t+h)$ zum Zeitpunkt t+h mit Hilfe von vier Zwischenauswertungen aus den Werten für $P_j(t)$ zum Zeitpunkt t berechnet:

$$P_j(t+h) = P_j(t) + \frac{h}{6} \cdot [k_{1j} + 2k_{2j} + 2k_{3j} + k_{4j}] \quad , \quad j = 0,\ldots,N. \qquad (4.54)$$

Dabei ist h die Integrationsschrittweite. Die Runge-Kutta (RK) Koeffizienten $k_{1j}, k_{2j}, k_{3j}, k_{4j}$ erhält man durch Auswertung des Differentialgleichungssystems mit den zur Zeit t_x gültigen Werten für die Übergangsraten und mit den Werten $Z_j(t_x)$ für die Zustandswahrscheinlichkeiten:

$$k_{mj} = -q_j(t_x) \cdot Z_j(t_x) + \sum_{k \neq j} q_{kj}(t_x) \cdot Z_k(t_x) \quad , \quad m = 1,\ldots,4 \qquad (4.55a)$$

$$
\begin{aligned}
k_{1j} &: t_x = t && , Z_j(t_x) = P_j(t) \\
k_{2j} &: t_x = t + h/2 && , Z_j(t_x) = P_j(t) + k_{1j} \cdot h/2 \\
k_{3j} &: t_x = t + h/2 && , Z_j(t_x) = P_j(t) + k_{2j} \cdot h/2 \\
k_{4j} &: t_x = t + h && , Z_j(t_x) = P_j(t) + k_{3j} \cdot h \quad .
\end{aligned}
\tag{4.55b}
$$

Da beim Runge-Kutta-Verfahren 4. Ordnung die RK-Koeffizienten nur einmal pro Integrationsschritt benötigt werden, kann durch eine sofortige, gewichtete Summenbildung Speicherplatz eingespart werden. In dieser Version ist der Speicherbedarf für die Berechnung transienter Zustandswahrscheinlichkeiten nur viermal größer als der Zustandsraum: alte Zustandswahrscheinlichkeiten, neue Zustandswahrscheinlichkeiten oder Zwischenwerte, summierte RK-Koeffizienten, momentane RK-Koeffizienten. Zur Verbesserung der Rechenzeit, ohne jegliche Einbuße der Genauigkeit, wurde die Schrittweite h während des Rechenvorganges adaptiv eingestellt. Dazu wurde ein zeitabhängiger Faktor $D_{max}(t+h)$ als Maß für die Systemdynamik zur Zeit t+h eingeführt. Diesem Wert entspricht der größte Differentialquotient im betrachteten Zeitpunkt

$$
D_{max}(t+h) = \max_j \left| P_j(t+h) \right| \tag{4.56}
$$

Im Falle einer großen Systemdynamik weicht $D_{max}(t+h)$ stark vom vorherigen Wert $D_{max}(t)$ ab, und je nach gewähltem Kriterium muß der Runge-Kutta-Schritt mit kleinerer Schrittweite wiederholt werden. Bei kleinen Differenzen - Annäherung an einen stationären Zustand - kann die Schrittweite entsprechend vergrößert werden. Der komplette Algorithmus ist in [van As (1984a)] enthalten.

4.5 Verkehrssimulation

Unter Verkehrssimulation versteht man die Abbildung des dynamischen Ablaufgeschehens von Verkehrsmodellen mit Hilfe eines Programmes auf einem Digitalrechner.

Im wesentlichen wird die Simulationstechnik angewendet:

- für neue Probleme, deren analytischer Lösungsweg noch unbekannt ist,
- zur Überprüfung von Modellierungsannahmen (Abhängigkeiten, Prozeßapproximationen),
- zur Validierung von approximativ berechneten, analytischen Ergebnissen,
- und für die Untersuchung von Modellen mit engen wechselseitigen Beziehungen zwischen verschiedenen Parametern, wie dies beispielsweise bei Kommunikationsprotokollen, Datenflußsteuerungen und Überlastabwehrstrategien in Paketvermittlungsnetzen der Fall ist.

Neben problemspezifischen Methoden, die auf besonderen Eigenschaften des Verkehrsmodells basieren - Roulette Simulation [Kosten (1970), van As (1975)], Teilruf-Simulation [Dietrich/Salade (1977)], Regenerative Simulation [Iglehart/Shedler (1980)]- wird die zeittreue, ereignisgesteuerte Simulation (time-true simulation oder event-by-event simulation) mit Erfolg für simulative Untersuchungen von Verkehrsmodellen angewendet [Fishman (1973, 1978), Gross/Harris (1974), Kühn (1981), Law/Kelton (1982)].

4.5.1 Allgemeine Prinzipien

Bei der zeittreuen, ereignisgesteuerten Simulation wird das Ablaufgeschehen im Modell bestimmt durch eine Folge von Ereignissen, die zu diskreten Zeitpunkten Zustandsänderungen bewirken. Programmtechnisch wird dies verwirklicht durch die Führung einer zeitlich geordneten Ereignisliste, deren Ereignisse je nach Typ (Ankunft, Bedienungsende, Zeitmarke) mit einer Reihe von entsprechenden Aktivitäten verbunden sind: Belegen einer Bedienungseinheit, Einreihen in die Warteschlange, Planen eines zukünftigen Ereignisses, usw. Alle Aktivitäten, die zu einem Ereignis gehören, werden zum Ereigniszeitpunkt ausgeführt. Falls, in Ausnahmefällen, die Zeit zur Ausführung dieser Aktivitäten (z.B. Verwaltungsarbeit) ebenfalls simulativ zu berücksichtigen ist, kann dies durch ein zusätzliches Ereignis

geschehen. Auf diese Weise ist das Verkehrsgeschehen auf die Ereigniszeitpunkte konzentriert und ohne Informationsverlust kann von einem Ereignis zum nächsten gesprungen werden, so daß im allgemeinen eine Zeitraffung möglich ist. Deshalb ist die Anzahl der simulierten Ereignisse und deren Behandlungsaufwand maßgebend für die Laufzeit eines Simulationsprogramms.

Struktur und Betriebsorganisation des Verkehrsmodells, wie zum Beispiel in Bild 4.1 für ein einstufiges Warteschlangenmodell, werden reflektiert in einer entsprechenden Daten- und Programmstruktur. Stochastische Verkehrsparameter (Ankunftsabstände, Bedienungszeiten, Gruppengrößen) werden mit Hilfe von gleichverteilten Pseudo-Zufallszahlen gemäß einer vorgegebenen Verteilungsfunktion erzeugt. Vervollständigt wird das Simulationsmodell durch Messung von Zeiten, Zählen von Ereignissen und die statistische Auswertung, die je nach Simulationsmodus - stationär oder transient - teilweise auf verschiedene Weisen durchgeführt werden.

4.5.2 Die stationäre Simulation

Charakteristisch für die stationäre Simulation ist die Unterteilung des Simulationslaufes in einen Vorlauf zur Erreichung des stationären Zustandes und in N Teiltests zur Ermittlung einer statistischen Aussagesicherheit der gemessenen Verkehrsgrößen.

Während der Simulation werden Meßdaten gesammelt, die am Ende jedes Teiltests ausgewertet werden. Auf diese Weise erhält man für jede ermittelte Verkehrsgröße N Stichproben, aus denen ein Mittelwert und mit Hilfe des Student-t-Tests ein Vertrauensintervall für eine vorgegebene Aussagesicherheit (z.B. 95%) gewonnen werden kann. Als Richtwert gilt: N = 10 mit je 10000 Ankunftsereignissen.

4.5.3 Die transiente Simulation

Ausgangspunkt für die transiente Simulation ist ein leeres System oder ein anderer genau definierter Anfangszustand. Mit diesem Beginnzustand wird während eines vorgegebenen Zeitintervalls simuliert, und es werden an vorher festgelegten Meßzeitpunkten Meßdaten gesammelt. Dies wird als Elementartest bezeichnet.

Analog zum stationären Fall werden die Meßwerte und die zugehörigen Vertrauensintervalle ermittelt, indem der Simulationslauf in N Teiltests unterteilt wird, die je eine ausreichend große Anzahl M von Elementartests enthalten (Richtwert: N = 10, M = 500).

Zur Messung von Warte- oder Durchlaufzeiten werden an den Meßzeitpunkten virtuelle Testanforderungen erzeugt, deren Schicksal verfolgt wird. Diese Testanforderungen dürfen das Verkehrsgeschehen nicht beeinträchtigen, denn sie sind nur virtuell vorhanden, d.h. sie beanspruchen keine Betriebsmittel. Ferner muß, über das Simulationsintervall hinaus, so lange simuliert werden, bis die Warte- oder Durchlaufzeit aller Testanforderungen des momentanen Elementartests bestimmbar ist.

Außer in der Organisation und im Meßverfahren zeichnet sich die transiente oder instationäre Simulation dadurch aus , daß zeitabhängige Ereignisgeneratoren für den Ankunfts- bzw. Bedienungsprozeß erforderlich sind [Tran-Gia (1982, 1983)].

4.6 Programmsysteme zur Untersuchung von Verkehrsmodellen

Der zum Teil beträchtliche Arbeitsaufwand um ein Verkehrsmodell in ein lauffähiges Programm umzusetzen, kann durch eine flexible und wirksame Software-Unterstützung wesentlich verringert werden. In diesem Kapitel werden die Merkmale von zwei Programmsystemen, die im Rahmen dieser Arbeit entwickelt wurden, kurz beschrieben.

Aufgrund dieser Software-Unterstützung können

- die Routineaufgaben weitgehend vermieden werden,
- die Programmentwicklungszeiten entscheidend verkürzt werden,
- die kreativen Tätigkeiten ausschließlich auf die modellspezifischen Programmteile konzentriert werden,
- die Softwarefehler größtenteils auf diese neuen Programmteile beschränkt werden.

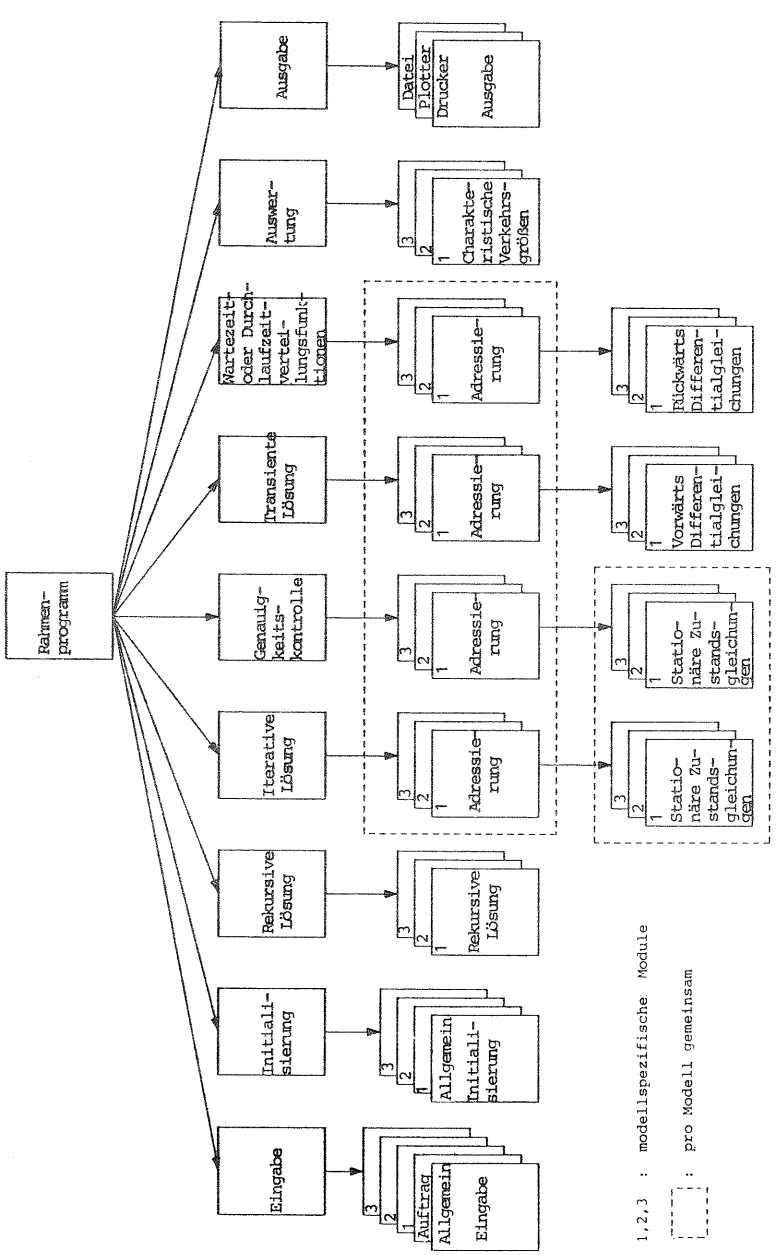


Bild 4.4: Struktur eines Programmsystems zur allgemeinen numerischen Lösung von Markoff-Warteschlangenmodellen.

4.6.1 Programmsystem zur numerischen Lösung von Markoff-Modellen

Dieses Programmsystem ist ein universelles Hilfsmittel zur analytischen Berechnung des stationären und transienten Ablaufgeschehens in komplexen mehrdimensionalen Markoff-Modellen. Es enthält die Softwaremodule für die stets wiederkehrenden Aufgaben und es legt den organisatorischen Rahmen für das Einfügen von modellspezifischen Programmteilen fest. Während der Entwicklung wurde auf eine flexible Modularität, eine große Transparenz und eine leichte Erweiterbarkeit geachtet.

Der Kern dieses Programmsystems bildet die Berechnung der

- stationären Zustandswahrscheinlichkeiten P_j, $j = 0,\ldots,N$ mit Hilfe der Kolmogoroff-Vorwärts-Gleichungen (4.19),
- transienten Zustandswahrscheinlichkeiten $P_j(t)$, $j = 0,\ldots,N$ mit Hilfe der Kolmogoroff-Vorwärts-Differentialgleichungen (4.16),
- bedingten komplementären Wartezeit- bzw. Durchlaufzeitverteilungsfunktionen $f_i(s,t)$, $i = 0,\ldots,N$ mit Hilfe der Kolmogoroff-Rückwärts-Differentialgleichungen (4.23).

Aus diesen primären Ergebnissen werden die charakteristischen Verkehrsgrößen des betrachteten Markoff-Modells berechnet. Das Programmsystem benötigt für jeden Zustand des Markoff-Modells eine explizite Darstellung der zugehörigen Gleichung bzw. Differentialgleichung. Programmtechnisch werden alle Zustände mit demselben Gleichungstyp zusammengefaßt. Je nach Komplexität benutzt man entweder eine allgemeine Gleichung (oder Differentialgleichung) mit sämtlichen Übergängen und Übergangswahrscheinlichkeiten gemäß dem Gleichungstyp oder eine Gleichung für jede Zustandsgruppe.

Um modellunabhängige Algorithmen (SOR, Runge-Kutta) zu erhalten, werden die im allgemeinen mehrdimensionalen Zustandsräume durch eine modellspezifische Adressierung in einen eindimensionalen Zustandsraum umgewandelt.

Die Struktur dieses Programmsystems zeigt Bild 4.4. Die höchste Programmebene besteht aus einem Rahmenprogramm, das im Verlauf

des Rechenvorganges bis zu 9 Steuerungsmodule aufrufen kann:

- die Eingabe mit einer Verzweigung in Module für die Eingabe von allgemeinen Daten, für die Eingabe von modellspezifischen Daten und für die Eingabe der Auswertungsauftragsliste. Diese Auftragsliste bestimmt, welche Verkehrsmodelle ausgewertet werden sollen, welche Ergebnisse zu berechnen sind und in welcher Reihenfolge die einzelnen Steuerungsmodule und deren Verzweigungen vom Rahmenprogramm aufzurufen sind.
- die Initialisierung aufgeteilt in allgemeine Initialisierung und modellspezifische Initialisierung.
- die rekursive Lösung zum Aufruf von modellspezifischen rekursiven Lösungsalgorithmen.
- die iterative Lösung zur Ausführung der SOR-Punktiteration gemäß Abschnitt 4.4.1 unter Verwendung der modellspezifischen Zustandsgleichungen und deren Adressierung. Bei Parameterstudien wird die zuletzt berechnete Zustandsverteilung als Startverteilung für den nächsten Parameter-Satz benutzt.
- die Genauigkeitskontrolle zur Überprüfung der Genauigkeit der rekursiv berechneten Zustandswahrscheinlichkeiten (siehe Abschnitt 4.4.2). Falls die Genauigkeit dem geforderten Wert nicht entspricht, wird nachiteriert.
- die transiente Lösung zur Ausführung des Runge-Kutta-Verfahrens nach Abschnitt 4.4.3 unter Verwendung der modellspezifischen Differentialgleichungen für die Zustände und deren Adressierung.
- die Wartezeit- oder Durchlaufzeitverteilungsfunktionen zur Ausführung des Runge-Kutta-Verfahrens nach Abschnitt 4.4.3 unter Verwendung der modellspezifischen Rückwärts-Differentialgleichungen für die bedingten komplementären Verteilungsfunktionen und deren Adressierung. Die Verteilungsfunktion selbst wird nach Abschnitt 4.2.4 durch eine entsprechende Gewichtung mit den Antreffwahrscheinlichkeiten berechnet.
- die Auswertung zum Aufruf von modellspezifischen Modulen für die Berechnung charakteristischer Verkehrsgrößen aus den Zustandswahrscheinlichkeiten oder zur numerischen Integration der Wartezeit- oder Durchlaufzeitverteilungsfunktionen.

- die Ausgabe zur Festlegung der Ausgabeform oder zum Abspeichern der Resultate.

4.6.2 Programmsystem für simulative Untersuchungen

Die zahlreichen und komplexen Verkehrsmodelle zur Untersuchung von Paketvermittlungsnetzen und Kommunikationsprotokollen erfordern eine flexible und wirksame Software-Unterstützung zur Erstellung der Simulationsprogramme. Gefordert wird eine übersichtliche und schnellere Programmerstellung unter Berücksichtigung von Rechenzeit- und Speicherplatzökonomie. Zu diesem Zweck gibt es heute schon zahlreiche Simulationssprachen, von denen insbesondere SIMULA [Rohlfing (1973), Lamprecht (1976)], GPSS [Rösmann (1978), Gordon (1978)], GPSS-FORTRAN [Schmidt (1980)] und SIMSCRIPT [Kampe (1971)] verwendet werden.

Nachteile von Simulationssprachen sind aber, daß

- sie nicht so allgemein zur Verfügung stehen wie zum Beispiel die universelle Programmiersprache FORTRAN,
- sie nicht kostenlos benutzt werden können,
- sie oft ohne Quellencode vom Hersteller geliefert werden, so daß weder Anpassungen vorgenommen werden können, noch die implementierten Aktionen kontrolliert werden können,
- sie oft für allgemeine Simulationszwecke gestaltet sind, so daß mehr Speicherplatz oder Rechenzeit benötigt wird als für die Simulation von Verkehrsmodellen notwendig wäre,
- sie nicht immer die gewünschte Flexibilität aufweisen.

Aus diesen Gründen wurde eine Bibliothek von Software-Modulen in FORTRAN IV aufgebaut, in der die meisten Komponenten einer Verkehrssimulation enthalten sind. Ergänzt durch eine flexible Datenstruktur bietet dieses Simulationssystem eine hervorragende Unterstützung zur Erstellung von Programmen für stationäre und transiente Verkehrssimulationen unter Berücksichtigung von anwenderspezifischen Forderungen. Das als QSIMLIB (Queueing Systems SIMulation LIBrary) bezeichnete Simulationssystem [Dehl (1981), van As (1984b)], wurde an Hand von vielen Verkehrsmodellen in dieser Arbeit und in einer Reihe von weiteren

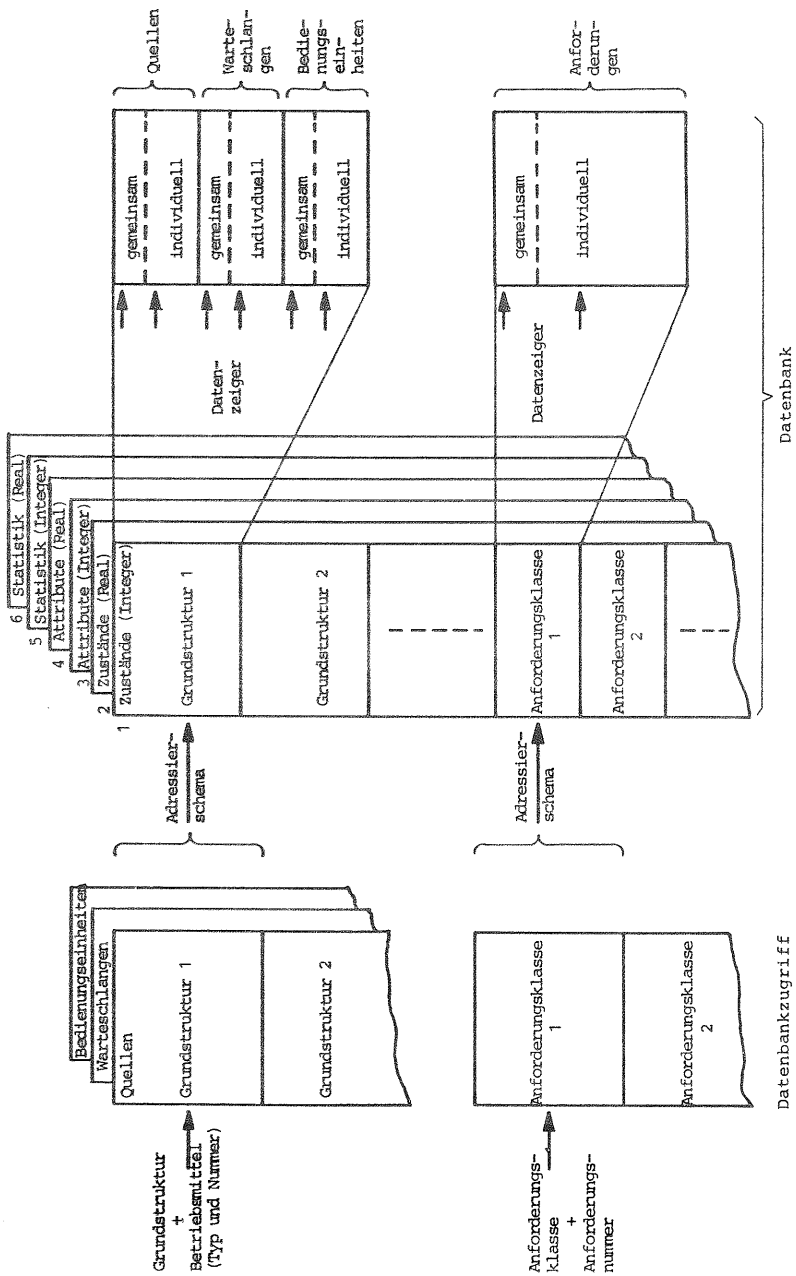


Bild 4.5: Simulationssystem QSIMLIB: Datenbankstruktur und Datenbankzugriff.

Anwendungen erprobt [Schwanke (1981), Schäfer (1982), Kiesel/ Kühn (1982), van As (1982), Manfield/Tran-Gia (1983), Finger (1983), Fischer (1983), Vogt (1983)]. Charakterisiert wird das Simulationssystem durch einige Merkmale, die auf den folgenden Seiten beschrieben sind.

4.6.2.1 Einstufiges Warteschlangensystem als Grundstruktur

Als Grundstruktur wird ein einstufiges Warteschlangensystem (vgl. Bild 4.1), bestehend aus Quellen, Warteschlangen und Bedienungseinheiten verwendet. Zu jeder Grundstruktur gehören Daten und ein Software-Modul zur Beschreibung des internen Ablaufgeschehens. Dieser Software-Modul ist gemeinsam für alle Grundstrukturen mit gleichem Verkehrsablauf. Identische Warteschlangensysteme eines Netzmodells können wahlweise als einzelne Grundstrukturen deklariert werden oder in einer einzigen Grundstruktur zusammengefaßt werden.

4.6.2.2 Anforderungsklassen

Die Anforderungen, die das Verkehrsmodell durchlaufen, werden eingeteilt in Anforderungsklassen. In der Regel werden verschiedene Klassen nur verwendet, wenn Anforderungsgruppen mit unterschiedlichem Datenbedarf pro Anforderung existieren.

Als Beispiel kann hier die detaillierte Simulation einer virtuellen Verbindung unter Einfluß des Netzverkehrs angeführt werden. Pakete, die zur virtuellen Verbindung gehören, müssen eine Reihe von Daten mitführen, während Pakete der Netzquellen (Hintergrundverkehr) nur wenige Daten brauchen.

4.6.2.3 Datenbankstruktur

Zur modellunabhängigen Verwaltung der Daten wird eine Datenbankstruktur verwendet. Diese Datenbank organisiert sich selbst aufgrund von Eingabedaten für das betreffende Verkehrsmodell. Datenbank und Datenbankzugriff sind in Bild 4.5 schematisch dargestellt.

Grundsätzlich werden die Daten für Zustände, Attribute und Statistik in getrennten Datenbereichen abgespeichert. Bedingt

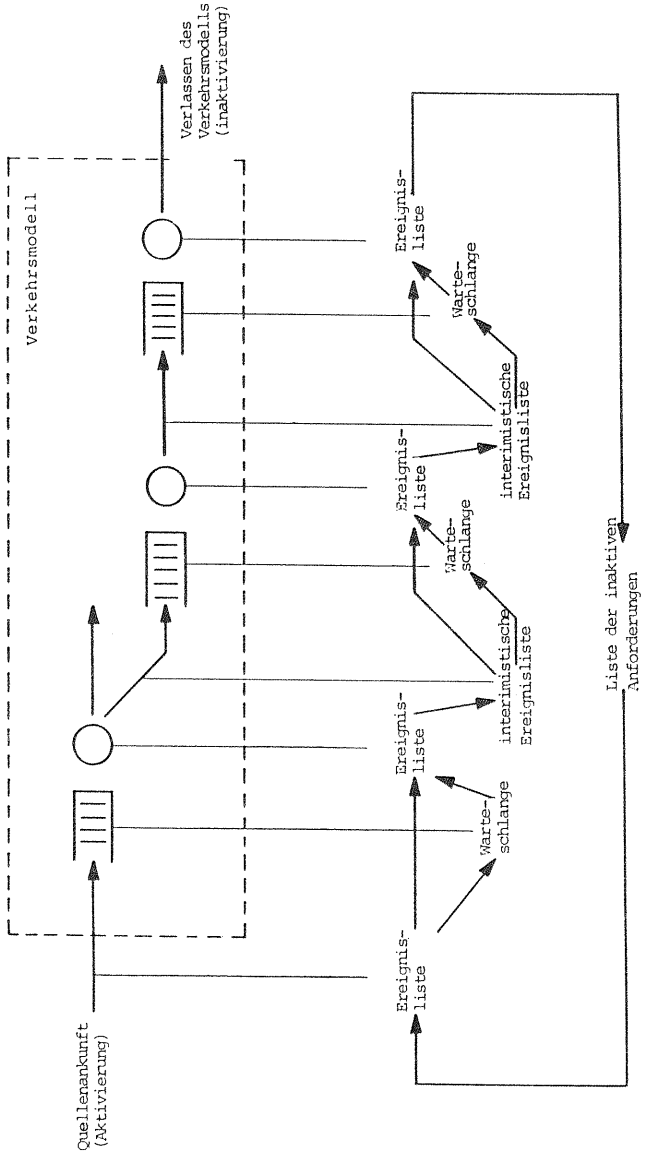


Bild 4.6: Simulationssystem QSIMLIB: Lebenslaufzyklus von Anforderungen.

durch die verwendete Sprache FORTRAN muß jeweils zwischen sogenannten Integer- und Realzahlen unterschieden werden, so daß es insgesamt sechs Datenbereiche gibt. Jeder dieser Datenbereiche ist initialisierungsbedingt aufgeteilt in Bereiche für die Grundstrukturen und die Anforderungsklassen. Innerhalb einer Grundstruktur wird weiter aufgeteilt in Daten für Quellen, Warteschlangen und Bedienungseinheiten, wobei unterschieden wird zwischen gemeinsamen und individuellen Daten. Diese Unterscheidung gilt auch für die Daten einer Anforderungsklasse. Zum Beispiel haben die Anforderungen derselben Klasse eine gemeinsame Statistik, aber jede Anforderung für sich hat individuelle Daten. Ferner kann beispielsweise die Warteschlangenkapazität als gemeinsame Angabe abgespeichert werden, wenn alle Warteschlangen in einer Grundstruktur dieselbe Kapazität besitzen, oder sie kann individuell für jede Warteschlange angegeben werden.

Ein Datenzugriff ist nur über ein hierarchisches Adressierschema möglich. Zum Beispiel müssen durch Angabe von Grundstruktur und Quellennummer sogenannte Datenzeiger gesetzt werden, bevor auf die Daten einer Quelle zugegriffen werden kann.

4.6.2.4 Lebenslaufzyklus

Die Anforderungen durchlaufen das Verkehrsmodell in einem sogenannten Lebenslaufzyklus, indem sie von einer Liste zur nächsten transferiert werden (Bild 4.6).

Es wird unterschieden zwischen den Listen der inaktiven Anforderungen, den Warteschlangen, der interimistischen Ereignisliste und der Ereignisliste selbst. Die Listen der inaktiven Anforderungen und die Warteschlangen sind einfach verkettete Listen, die stets sequentiell durchlaufen werden. Die Ereignisliste besteht aus einem binären Baum, während die interimistische Ereignisliste als einfacher, zyklischer Zwischenspeicher organisiert ist. Er dient dazu, Anforderungen bei einem Wechsel zwischen Grundstrukturen zwischenzuspeichern. Dadurch ist die Modularität in der Modellierung gewährleistet. Der Transfer von Anforderungen von Liste zu Liste geschieht lediglich durch Änderung der Verkettungsstruktur der Listen. Die eigentlichen Daten

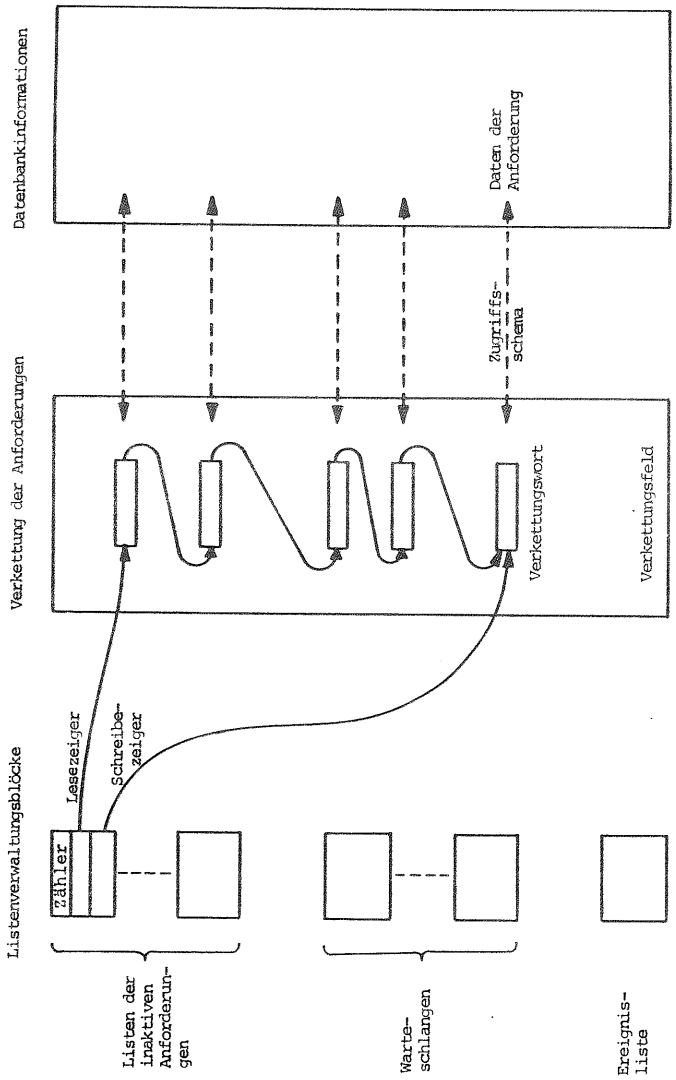


Bild 4.7: Simulationssystem QSIMLIB: Lineare Listenstruktur.

der Anforderungen selbst, die in der Datenbank abgespeichert sind, werden nicht transferiert.

4.6.2.5 Einfach verkettete Listen

Jede Anforderung, die vor Simulationsbeginn aufgrund der Eingabe generiert wird, besitzt einen fest zugeordneten Speicherplatz (Verkettungswort) in einem sogenannten Verkettungsfeld (Bild 4.7). Die Verkettungsworte dienen dazu, die Anforderungen, die zur gleichen einfach verketteten Liste gehören, miteinander zu verknüpfen. Die Listen der inaktiven Anforderungen einer Anforderungsklasse und die Warteschlangen haben diese Organisationsform. Sie werden sequentiell durchlaufen und benötigen dazu einen eigenen Listenverwaltungsblock. Er besteht aus einem Zähler (Anzahl der Anforderungen in der Liste), einem Lesezeiger (erste Anforderung in der Liste) und einem Schreibezeiger (letzte Anforderung in der Liste). Obwohl die Listen in der Regel in der Verkettungsreihenfolge verarbeitet werden, kann an jeder Stelle der Liste eine Anforderung entfernt (z.B. bei zufälliger Abfertigung einer Warteschlange) oder hinzugefügt werden.

Die Identität einer Anforderung ist durch die Position ihres Verkettungswortes innerhalb des Verkettungsfeldes bestimmt. Hieraus kann die Anforderungsklasse und die Anforderungsnummer ermittelt werden, so daß über den Datenbankzugriffsmechanismus die eigentlichen Daten der betreffenden Anforderung zugänglich sind.

4.6.2.6 Ereignisliste

Die Ereignisliste ist als binärer Baum organisiert, so daß viele Ereignisse mit minimalem Zeitaufwand verwaltet werden können (Bild 4.8). Jedes Element dieses Baumes besteht aus drei Zeigern (linker Nachfolger, rechter Nachfolger, Vorgänger), der Ereigniszeit und einer Ereignisidentität. Eine negative Ereignisidentität bedeutet, daß es sich um eine Zeitmarke handelt. In diesem Fall gibt die Ereignisidentität zugleich die Ereignisaktion an. Ist die Ereignisidentität positiv, so ist sie identisch mit der Anforderungsidentität. Aufgrund dieser

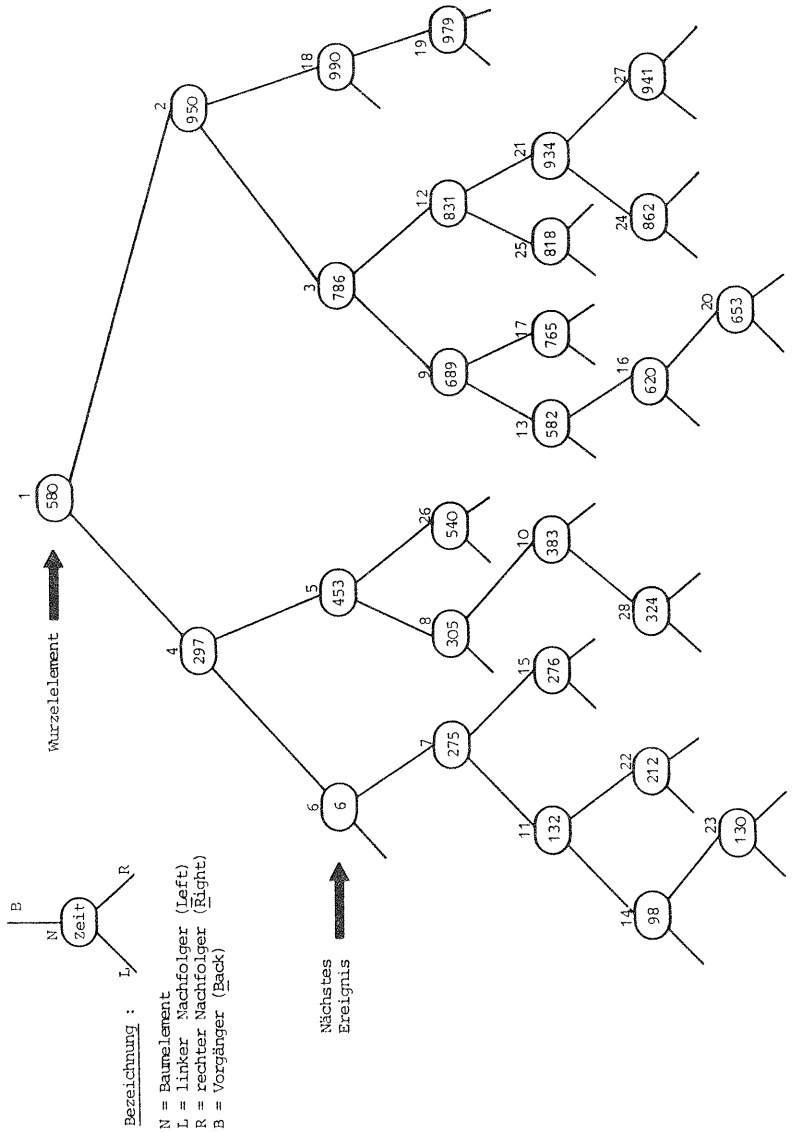


Bild 4.8: Simulationssystem QSIMLIB: Ereignisliste mit binärer Baumstruktur.

Identität wird auf die Daten der betreffenden Anforderung in der Datenbank zugegriffen. Sie enthalten u.a. Angaben über die Position der Anforderung im Verkehrsmodell und über die vom Ereignis ausgelösten Aktion.

Ferner benötigt der binäre Baum einen Zeiger, um den Baumanfang (Wurzelelement) anzugeben. Ein zweiter Zeiger identifiziert das nächste Ereignis. Eine vorgegebene Anzahl von Ereigniselementen teilt sich auf in aktive und nichtaktive Elemente. Die aktiven Elemente, deren Anzahl in einem Zähler festgehalten werden, bilden zusammen den binären Baum. Die momentan nichtaktiven Elemente sind als einfach verkettete Liste zusammengefügt.

Grundoperationen für den binären Ereignisbaum:

- Ein neues Ereignis wird mit Hilfe von wenigen Binärvergleichen am Ende des Baumes chronologisch einsortiert.
  Beispiel: Ein Ereigniselement mit Ereigniszeit 350 wird am rechten Nachfolgerzeiger vom Baumelement 28 verkettet.
- Bei Entfernung des nächsten Ereignisses muß der Baum aktualisiert und das nächste Ereignis gesucht werden.
  Beispiel: Bei Entfernung vom Baumelement 6 mit Ereigniszeit 6 muß Element 7 mit Element 4 verknüpft werden und der Zeiger für das nächste Ereignis muß auf Element 14 zeigen.
- Mitunter ist es notwendig, Ereignisse zu annullieren (unterbrechende Prioritäten, Löschen von Time-Out Ereignissen). Ereignisse mit keinem oder einem einzigen Nachfolger können direkt eliminiert werden. Ist dies nicht der Fall, so muß ein geeignetes Baumelement gesucht werden, dessen Entfernung die Baumstruktur nicht zerstört. Danach wird seine Ereignisinformation (Zeit, Identität) transferiert und dieses Baumelement anstelle des ursprünglichen Elementes eliminiert.
  Beispiele: Baumelement 28 kann direkt eliminiert werden; bei Entfernung von Baumelement 10 müssen die Elemente 28 und 8 miteinander verknüpft werden; bei Entfernung von Baumelement 3 wird der Inhalt von Element 25 nach Element 3 transferiert und anschließend Element 25 eliminiert.

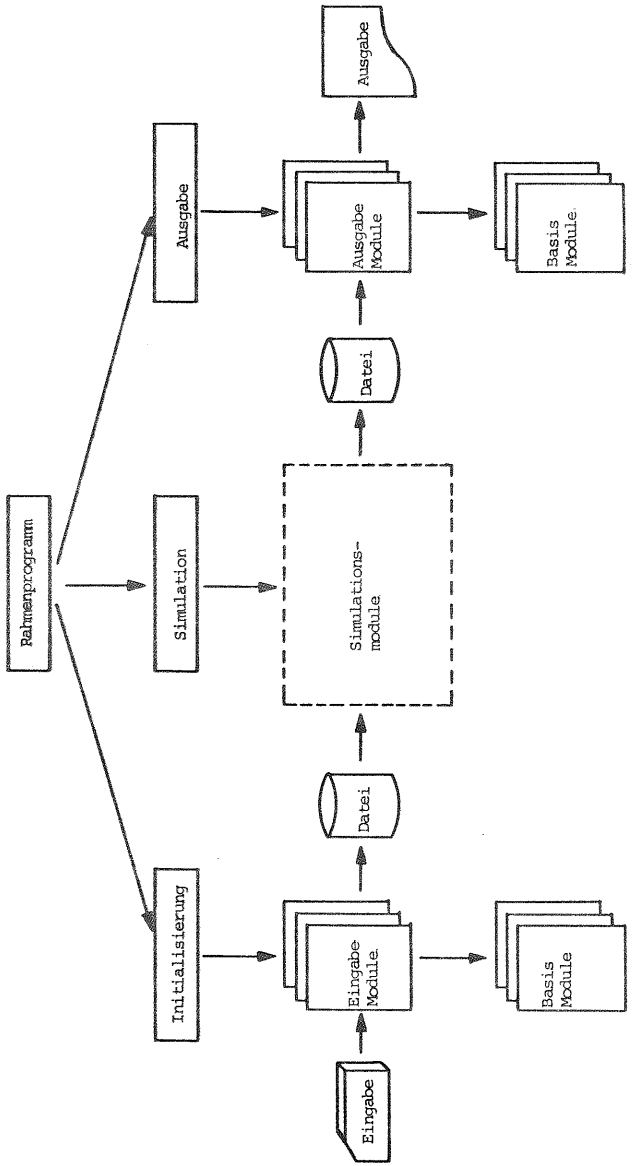


Bild 4.9: Simulationssystem QSIMLIB: Programmstruktur.

4.6.2.7 Programmstruktur

Das Programmsystem ist hierarchisch und modular aufgebaut (Bild 4.9). Es besteht aus drei unabhängigen Programmteilen (Initialisierung, Simulation, Ausgabe), die in der Regel von einem übergeordneten Rahmenprogramm nacheinander aufgerufen werden.

Bei Simulationsmodellen mit einem hohen Speicherbedarf ist es von Vorteil diese Programmteile getrennt zu benutzen. Die Daten werden dann über Dateien ausgetauscht. Mit Hilfe der Dateien kann der Simulationslauf selbst auch in mehrere Teile zerlegt werden. Bei transienten Simulationen wird die Datei dazu benutzt einen Anfangszustand abzuspeichern.

Alle drei Programmteile setzen sich zusammen aus modellspezifischen Modulen und Bibliotheksmodulen. In den Programmteilen für Initialisierung und Ausgabe sind "leere" Software-Module für die Implementierung von modellabhängigen Aktionen vorgesehen. Außer der Ablaufsteuerung wird der Simulationsteil dagegen völlig dem Modell angepaßt. Dies muß jedoch unter Berücksichtigung einiger Richtlinien und Vorschriften geschehen. Aufgrund der modularen Aufteilung in Grundstrukturen - dies betrifft sowohl die Modellstruktur als auch der Verkehrsablauf - können aber neue Verkehrsmodelle oft auf bereits bestehende Modelle aufgebaut werden.

Im Simulationsteil (Bild 4.10) gibt es zwei Gruppen von Modulen: einerseits die Module für den funktionellen Simulationsablauf und die Verkehrsmessung, andererseits die Module für die statistische Auswertung und für die Konsistenztests. Der Programmteil für den funktionellen Simulationsablauf setzt sich zusammen aus einzelnen Modulen, wobei jedes Modul den Verkehrsablauf einer Grundstruktur abbildet. Dabei ist jedes dieser Module aufgeteilt in vier Gruppen von Ereignisaktionen: Ankunft aus einer Quelle, Ankunft aus einer anderen Grundstruktur, Bedienungsende und Zeitmarke.

Die Ereignisaktionen enthalten in der Regel einige Bibliotheksmodule:

- Behandlung von Anforderungen (Zuteilungsstrategien, Abferti-

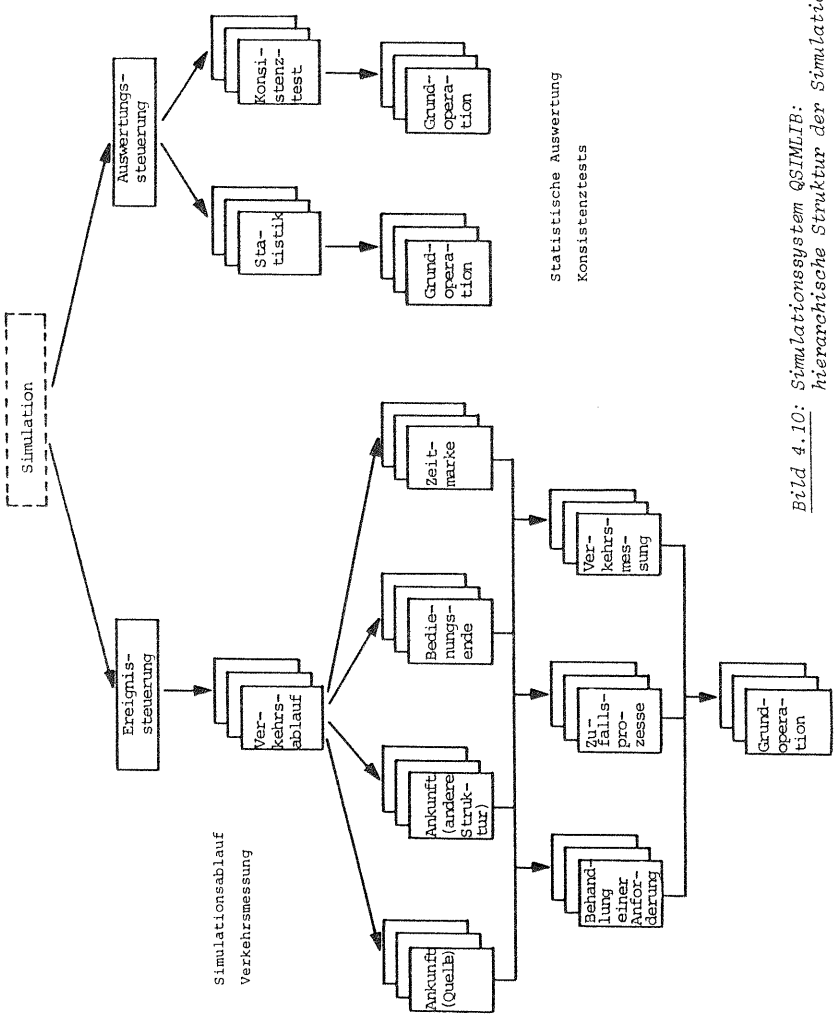


Bild 4.10: Simulationssystem QSIMLIB: hierarchische Struktur der Simulationsmodule

gungsstrategien, Prioritäten, Zutrittsberechtigung, Verkehrslenkung),

- Generierung von Zufallsgrößen (Ankunftsprozesse, Bedienungsprozesse, Gruppengröße),
- Verkehrsmessung (Zählung, Zeitmessung, Verteilungen),
- Grundoperationen (z.B. Einreihen in der Warteschlange, Belegen einer Bedienungseinheit, Adressierung für den Datenzugriff).

Für die Spezifikation der statistischen Auswertung sind wiederum "leere" Software-Module vorgesehen. Die Konsistenztests werden unterstützt, sie müssen jedoch für jedes Verkehrsmodell in einer anderen Form implementiert werden.

4.6.2.8 Simulationsablauf

In Übereinstimmung mit der Programmstruktur besteht der Simulationslauf aus einer Initialisierungs-, Simulations- und Ausgabephase. Der Ablauf der Simulationsphase richtet sich nach der Art der Simulation. Bei einer stationären Simulation besteht der Simulationslauf aus einer Vorlaufphase und einigen Teiltests. Bei transienten Simulationen besteht er aus Gruppen von Elementartests, die den Teiltest bilden. In beiden Fällen findet nach jedem Teiltest eine statistische Auswertung und eventuell ein Konsistenztest statt. Der Simulationsteil wird abgeschlossen mit einer statistischen Endauswertung.

Während der Simulation liefert die Ereignissteuerung stets das nächste Ereignis und je nach Typ wird in die entsprechende Ereignisverarbeitung verzweigt.

5. MODELLIERUNG VON TYPISCHEN ÜBERLASTSITUATIONEN

In diesem Kapitel betrachten wir Verkehrsmodelle, die das Verständnis für das Zustandekommen von typischen Überlastsituationen in Paketvermittlungsnetzen fördern sollen.

Im einzelnen werden behandelt:

- die Ausbreitung von Lastspitzen in Paketvermittlungsnetzen und deren Nachwirkung auf später eintreffende Pakete,
- der Rückstau von Paketen in einem Netzknoten infolge einer Überlastsituation in einem benachbarten Netzknoten,
- die lawinenartige Überflutung des Netzes mit Paketkopien bei Überschreitung einer kritischen Netzbelastung,
- der Einfluß von begrenzten Speicherkapazitäten auf Verkehrsströme mit sehr unterschiedlichen Verkehrsraten bei einer Abfertigung mit oder ohne Prioritäten.

5.1 Die Ausbreitung von Lastspitzen und deren Nachwirkung

Ausgangspunkt für unsere Betrachtung ist ein Verkehrsweg durch ein Paketvermittlungsnetz. Dieser wird modelliert durch eine Serie von Warteschlangensystemen, die die einzelnen Übertragungskanäle darstellen. Die Warteschlangensysteme sind vom Typ M/M/1.

5.1.1 Modellbeschreibung

Als Spezialfall eines Netzes wird ein N-stufiges Tandemmodell mit einer einzigen Verkehrsquelle betrachtet (Bild 5.1). Der nachstehende Berechnungsvorgang läßt sich jedoch auch auf allgemeine Warteschlangennetze, falls sie rückkopplungsfrei sind, anwenden.

Ausgehend von einer stationären Ankunftsrate wird in der Verkehrsquelle eine Überlastspitze erzeugt. Diese Überlastspitze entsteht durch Erhöhung der mittleren Ankunftsrate und hat die Form eines rechteckigen Impulses der Zeitdauer T und der

Intensität λ_{max}. Mit Hilfe dieses Verkehrsmodells wird nun die Ausbreitung dieser Lastspitze und ihre Nachwirkung auf später eintreffende Pakete quantitativ untersucht.

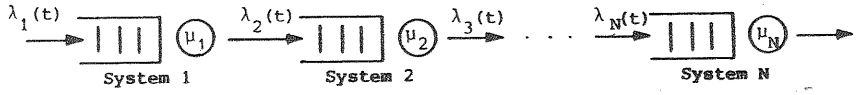


Bild 5.1: Verkehrsmodell: Ausbreitung einer Lastspitze und ihrer Nachwirkung.

5.1.2 Berechnungsverfahren

a) Voraussetzungen

Die Berechnung der Zustandswahrscheinlichkeiten beruht auf den folgenden Überlegungen und Annahmen:

- Nach dem Ausgangsprozeßtheorem von Burke [Burke (1956)] ist im stationären Fall der Ausgangsprozeß eines Wartesystems M/M/n wieder ein Poisson-Prozeß.
- Aufgrund dieser Tatsache kann das N-stufige Tandemmodell für den stationären Fall exakt zerlegt werden in die einzelnen M/M/1-Systeme. Dies geht auch aus der Produktlösungsform für Warteschlangennetze hervor.
- In [Upton/Tripathi (1982)] wird gezeigt, daß der transiente Ausgangsprozeß eines M(t)/M/1-Systems bei einer FIFO-Abfertigungsstrategie mit sehr guten Näherungen durch einen Poisson-Prozeß beschrieben werden kann. Die zugehörige Verteilungsfunktion für die Ausgangsabstände wird hierbei als eine gewichtete Funktion von zwei negativ exponentiellen Verteilungsfunktionen angenommen:

$$D(s,t) = Q(s,t) \cdot (1-e^{-\mu t}) + [1-Q(s,t)] \cdot (1-e^{-\lambda(t) \cdot t}) \quad . \qquad (5.1)$$

Dabei ist $Q(s,t) = 1$, wenn eine zum Zeitpunkt s eintreffende Anforderung das Warteschlangensystem zum Zeitpunkt t noch nicht verlassen hat, sonst gilt $Q(s,t) = 0$. Diese Gewichtungsfaktoren lassen sich berechnen durch Bestimmung der mittleren

Durchlaufzeit $t_F(s)$ nach Gl.(4.33) mit $S \to \infty$ bzw. Gl.(5.5).

- Bei einem hohen Verkehrsangebot und insbesondere in einer Überlastsituation wird der Ausgangsprozeß vorwiegend vom Bedienungsprozeß bestimmt.
- Bedingt durch die Näherungsannahme eines Poisson-Ausgangsprozesses können die M/M/1-Systeme auch für den transienten Fall unabhängig voneinander behandelt werden, wenn jeweils die für System i momentan gültige Ankunftsrate $\lambda_i(t)$ bekannt ist.

b) Zustandswahrscheinlichkeiten

Aufgrund der gemachten Annahmen können die transienten Zustandswahrscheinlichkeiten $P_{ij}(t)$, $i = 1,\ldots,N$ und $j = 0,1\ldots$, für das i-te Warteschlangensystem bei bekannter Ankunftsrate $\lambda_i(t)$ unabhängig von den anderen Systemen berechnet werden.

Für das i-te System lauten die Differentialgleichungen für die zeitabhängigen Zustandswahrscheinlichkeiten $P_{ij}(t)$

$$\frac{d}{dt} P_{i0}(t) = -\lambda_i(t) \cdot P_{i0(t)} + \mu_i \cdot P_{i1}(t)$$

$$\frac{d}{dt} P_{ij}(t) = -[\lambda_i(t)+\mu_i] \cdot P_{ij}(t) + \mu_i \cdot P_{i,j+1}(t) + \lambda_i(t) \cdot P_{i,j-1}(t) \quad , \tag{5.2}$$

$$j > 0 \quad , \quad i = 1,\ldots,N \quad .$$

Beginnend mit System 1 werden in jedem Runge-Kutta-Schritt die transienten Zustandswahrscheinlichkeiten der einzelnen Systeme sequentiell berechnet, wobei sich die momentane Ankunftsrate für das System (i+1) aus der momentanen Abgangsrate des Systems i bestimmen läßt:

$$\lambda_{i+1}(t) = [1-P_{i0}(t)] \cdot \mu_i \quad , \; i = 1,\ldots,N-1 \quad . \tag{5.3}$$

c) Mittlere Systembelastung

Aus den zeitabhängigen Zustandswahrscheinlichkeiten kann die mittlere Systembelastung zum Zeitpunkt t ermittelt werden:

$$E[X_i(t)] = \sum_{j=0}^{\infty} j \cdot P_{ij}(t) \quad , \quad i = 1,\ldots,N \quad . \tag{5.4}$$

d) Mittlere Durchlaufzeit

Unter Voraussetzung einer FIFO-Abfertigungsstrategie setzt sich die mittlere Durchlaufzeit in einem Warteschlangensystem einer Testanforderung mit Ankunftszeit s zusammen aus der Summe aller Bedienungszeiten der Anforderungen, die vorher eingetroffen waren, und der Bedienungszeit der Testanforderung selbst.

Wird somit beim Eintreffen am System i - zum Zeitpunkt s_i - eine mittlere Systembelastung $E[X_i(s_i)]$ vorgefunden, so gilt nach Gl.(4.33) für die mittlere Durchlaufzeit $t_{Fi}(s_i)$ in System i

$$t_{Fi}(s_i) = (E[X_i(s_i)]+1) \cdot h_i \quad , \quad h_i = \frac{1}{\mu_i} \quad , \quad i = 1,\ldots,N \quad . \tag{5.5}$$

Die gesamte Durchlaufzeit erhält man durch die zeitliche Verfolgung einer Testanforderung durch sämtliche Warteschlangensysteme. Dies ist in Bild 5.2 veranschaulicht.

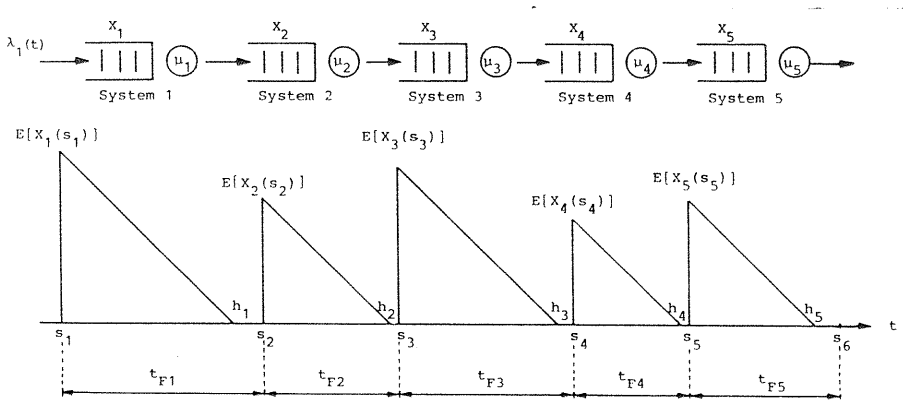


Bild 5.2: Zur Berechnung der Durchlaufzeit $t_F(s)$.

Die Anforderung trifft zum Zeitpunkt s_1 am System 1 ein. Mit der Beziehung (5.5) wird nun die Ankunftszeit s_2 am System 2 bestimmt. Aus der mittleren Systembelastung $E[X_2(s_2)]$ kann dann wiederum die Zeit s_3 berechnet werden, usw.

Für die gesamte mittlere Durchlaufzeit gilt somit:

$$t_F(s) = \sum_{i=1}^{N} t_{Fi}(s_i) \quad , \quad s = s_1 \quad . \tag{5.6}$$

Programmtechnisch wird die Bestimmung der Durchlaufzeit parallel zur Berechnung der zeitabhängigen Zustandswahrscheinlichkeiten durchgeführt. Für jede Testanforderung wird die Ankunftszeit s_i am nächsten System als Referenzzeit abgespeichert. Wird während des Berechnungsvorgangs für den Systemzustandsprozeß eine Referenzzeit erreicht, so kann bestimmt werden, wann die betreffende Testanforderung am nächsten System eintrifft oder wann das gesamte Tandemmodell durchlaufen ist. Es sei darauf hingewiesen, daß ähnlich wie bei der transienten Simulation, der Rechenprozeß erst beendet ist, wenn die Durchlaufzeit aller im betrachteten Zeitraum angenommenen Testanforderungen bestimmt worden ist. Der Systemzustandsprozeß muß deshalb über das Berechnungsintervall hinaus fortgesetzt werden.

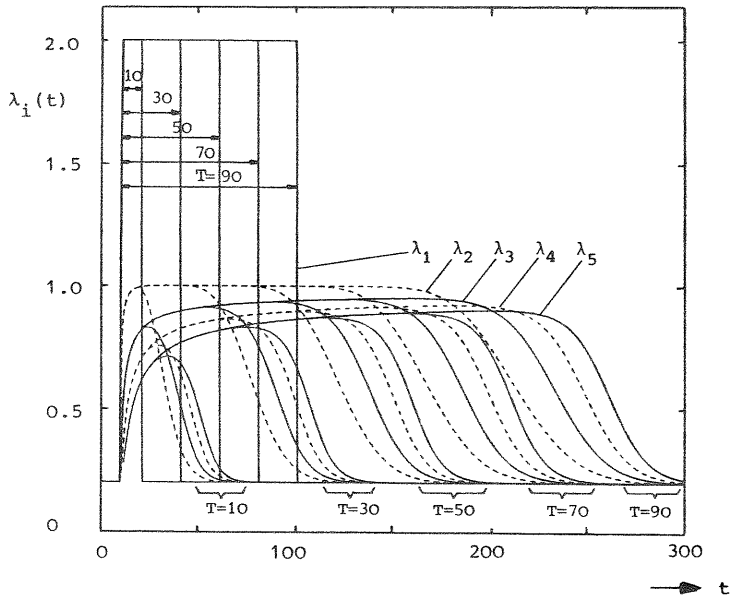
5.1.3 Numerische Ergebnisse

Als numerisches Beispiel wird ein 5-stufiges Tandem-System betrachtet. Die Bedienungsraten der einzelnen Systeme sind identisch: $\mu_1 = \mu_2 = \ldots = \mu_5 = 1$. Die Verkehrsquelle liefert eine stationäre Ankunftsrate $\lambda_1 = 0.2$ und eine rechteckförmige Überlastspitze mit einer Intensität $\lambda_{max} = 2$ und einer normalisierten Zeitdauer $T = 10,30,50,70,90$. In den nachfolgenden Bildern sind die Kurven der Systeme $i = 1,3,5$ durchgezogen, die der Systeme $i = 2,4$ gestrichelt.

a) Ankunftsrate

In Bild 5.3 ist die zeitabhängige Ankunftsrate $\lambda_i(t)$ dargestellt. Betrachtet man den schmalen Überlastimpuls $T = 10$, so sieht man, daß die Impulse der Ankunftsraten in den nachfolgenden Systemen kleiner, aber breiter werden. Außerdem werden

die Flanken immer flacher. Die Ankunftsrate im System 2 erreicht beinahe sofort ihren maximalen Wert $\lambda_{2max} = 1$, während die nachfolgenden Systeme 3,4 und 5 diese maximale Überlastrate erst bei länger andauernder Überlastsituation zu spüren bekommen. Eine hohe Überlastspitze am Eingang einer Serie von Warteschlangensystemen verwandelt sich also in einen wesentlich niedrigeren aber dafür viel breiteren Überlastimpuls. Die Fläche unterhalb der entsprechenden Überlastspitzen ist jeweils gleich groß und entspricht der Anzahl zusätzlich eintreffender Anforderungen. Bedingt durch die Vermaschung im allgemeinen Netz kumulieren sich die verschiedenen Ankunftsraten in den einzelnen Systemen. Somit können an jeder Stelle im Netz hohe Überlastspitzen entstehen.



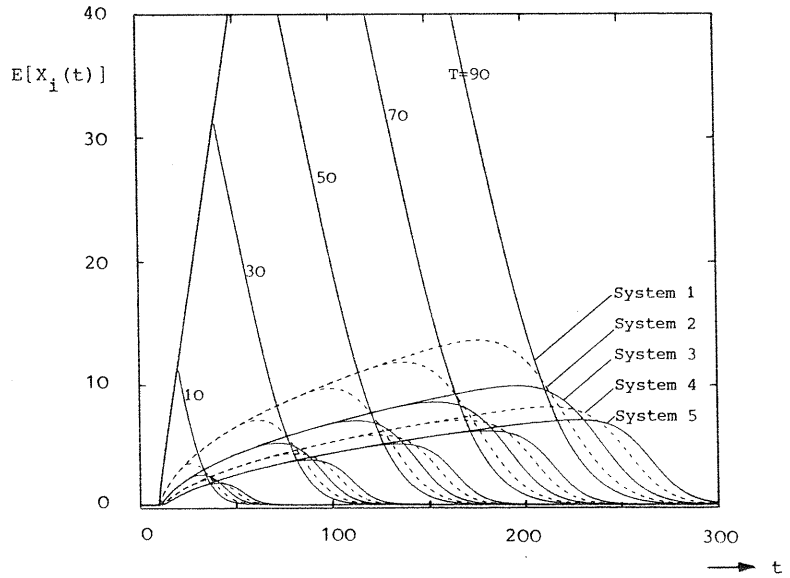
Ausbreitung eines Überlastimpulses in einem 5-stufigen Tandemmodell.

Bild 5.3: Ankunftsrate $\lambda_i(t)$ im Warteschlangensystem i bei verschiedenen Überlastimpulsdauern $T = 10,30,50,70,90$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 2.0$, $h_1 = \ldots = h_5 = 1$.

b) mittlere Systembelastung

Bild 5.4 zeigt die Systembelastung $E[X_i(t)]$ in den verschiedenen Warteschlangensystemen. Entsprechend den Ankunftsraten nimmt die Belastung im ersten System sehr schnell zu, während sie sich in den nachfolgenden Systemen wesentlich langsamer aufbaut.

Diese Ergebnisse lassen sich anschaulich sehr gut erklären: Mit dem Eintreffen des Überlastimpulses (z.B. für den Fall T=30) wird die Systembelastung im System 1 praktisch mit der Differenzrate $(\lambda_{max}-\mu_1)$ aufgebaut. Nach Beendigung des Überlastimpulses ist die mittlere Systembelastung $E[X_1(t)]$ etwa $(\lambda_{max}-\mu_1)\cdot 30 = 30$. Der Abbau danach erfolgt in erster Näherung mit der konstanten Rate μ_1 und ist dann etwa zur Zeit $t = 10+30+30 = 70$ im wesentlichen abgeschlossen.

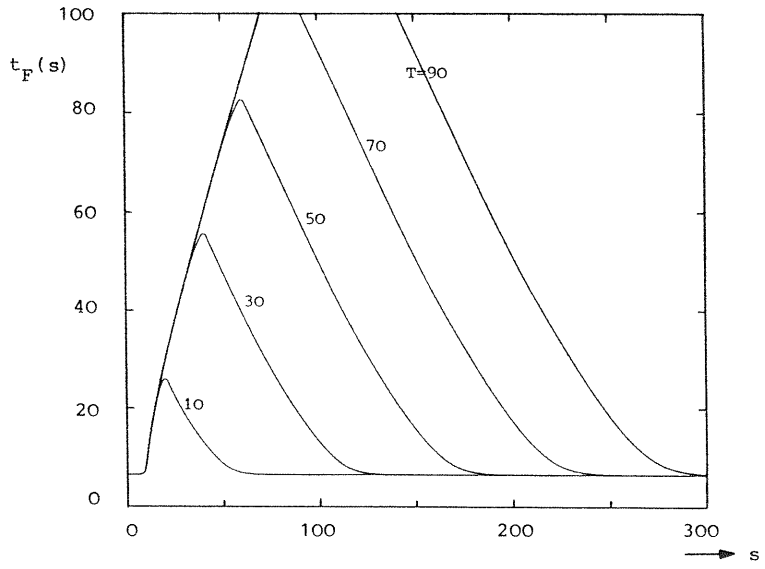


Ausbreitung eines Überlastimpulses in einem 5-stufigen Tandemmodell.
Bild 5.4: Mittlere Systembelastung $E[X_i(t)]$ im Warteschlangensystem i bei verschiedenen Überlastimpulsdauern $T = 10,30,50,70,90$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 2.0$, $h_1 = \ldots = h_5 = 1$.

Da die Abgangsrate des ersten Systems nie den Wert $\mu_1 = 1$ übersteigen kann, wirkt System 1 wie ein Lastspitzen-Filter für System 2. Die Zunahme der Systembelastung im zweiten System ist deshalb bedeutend geringer, da in Überlastfällen das System 2 ungefähr mit $\lambda_2 = 1$ belastet wird und gleichzeitig mit $\mu_2 = 1$ abgefertigt wird. Für die nachfolgenden Systeme 3, 4 und 5 gelten entsprechende Überlegungen.

c) mittlere Durchlaufzeit

Aus Bild 5.5 kann die mittlere Durchlaufzeit $t_F(s)$ von Anforderungen in Abhängigkeit von deren Ankunftszeit s entnommen werden. Aus den gezeigten Kurven läßt sich die Nachwirkung von Überlastspitzen deutlich erkennen: auch Anforderungen, die erst später eintreffen, müssen noch mit beträchtlichen Durchlaufzeiten rechnen.



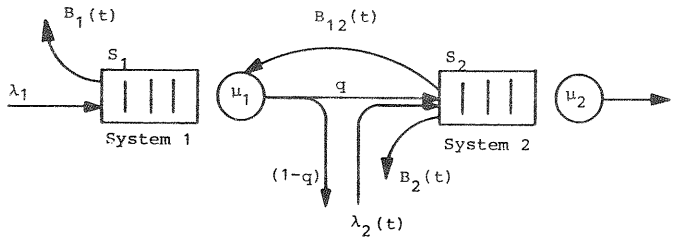
Ausbreitung eines Überlastimpulses in einem 5-stufigen Tandemmodell.
Bild 5.5: Mittlere Durchlaufzeit $t_F(s)$ als Funktion vom Ankunftszeitpunkt s bei verschiedenen Überlastimpulsdauern $T = 10,30,50,70,90$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 2.0$, $h_1 = \ldots = h_5 = 1$.

5.2 Der Rückstau von Paketen

Als nächstes wird die Entstehung eines Rückstaus, verursacht durch eine Überlastsituation im Nachbarknoten, gezeigt. Betrachtet wird jeweils der Ausgangsspeicher vor einem Übertragungskanal in zwei benachbarten Netzknoten. Ein Teil der Pakete wird auf ihrem Weg durch das Paketvermittlungsnetz von beiden Übertragungskanälen übertragen. Infolge der endlichen Größe der Speicher hat eine Überlastung des zweiten Übertragungskanals durch weitere Verkehrsströme eine direkte Rückwirkung auf die Verkehrseigenschaften des ersten Übertragungskanals, indem die momentan nicht akzeptierbaren Pakete eine erneute Übertragung benötigen. Es wird ferner angenommen, daß die entsprechenden Übertragungsprotokolle so schnell reagieren, daß die zugehörigen Verzögerungen vernachlässigbar sind.

5.2.1 Modellbeschreibung

Das entsprechende Verkehrsmodell ist in Bild 5.6 abgebildet und besteht aus zwei hintereinander angeordneten Warteschlangensystemen mit begrenzter Speicherkapazität. Im System 1 treffen Anforderungen nach einem Poisson-Prozeß mit zeitunabhängiger Ankunftsrate λ_1 ein. Im Zustand des vollen Systems, der mit der Verlustwahrscheinlichkeit $B_1(t)$ auftritt, werden diese Anforderungen abgewiesen und gehen verloren. Die akzeptierten Anforderungen werden in der Ankunftsreihenfolge (FIFO) bedient.



λ_1 = Ankunftsrate für System 1
$\lambda_2(t)$ = Ankunftsrate für System 2
$B_1(t)$ = Verlustwahrscheinlichkeit (System 1)
$B_2(t)$ = Verlustwahrscheinlichkeit (System 2)
$B_{12}(t)$ = Blockierungswahrscheinlichkeit (System 1 nach System 2)
q = Verzweigungswahrscheinlichkeit
μ_1 ,μ_2 = Bedienungsraten
S_1 ,S_2 = Systemkapazitäten

Bild 5.6: Verkehrsmodell: Entstehung eines Rückstaus von Paketen.

Nach Beendigung der negativ exponentiell verteilten Bedienungszeit mit Mittelwert $h_1 = 1/\mu_1$ im System 1 wird das Verkehrsmodell mit der Wahrscheinlichkeit 1-q verlassen, und mit der Wahrscheinlichkeit q werden die Anforderungen dem System 2 angeboten.

Für den Fall, daß das System 2 gerade voll ist wenn eine Anforderung vom System 1 eintrifft, - Blockierungswahrscheinlichkeit $B_{12}(t)$ - erfolgt eine erneute Bedienung durch System 1. Dies wiederholt sich solange bis System 2 die Anforderungen aufnehmen kann. Während der Wiederholungen ist der gesamte Verkehrsablauf im System 1 blockiert. Dem System 2 werden, außer den Anforderungen von System 1, auch weitere Anforderungen aus einem Poisson-Ankunftsprozeß mit zeitabhängiger Ankunftsrate $\lambda_2(t)$ angeboten. Diese zweite Verkehrsquelle dient dazu, die Überlastsituation im System 2 zu modellieren. Wenn Anforderungen aus dieser Quelle ein volles System 2 vorfinden, werden sie mit der Verlustwahrscheinlichkeit $B_2(t)$ abgewiesen.

5.2.2 Berechnungsverfahren

Bedingt durch die Blockierung zwischen beiden Warteschlangensystemen, kann das Verkehrsmodell nicht in Einzelsysteme zerlegt werden. Der Berechnung dieses Modells liegt deshalb ein zweidimensionales Markoff-Zustandsdiagramm zugrunde.

a) Zustandsdiagramm

Bild 5.7 zeigt die Struktur des Zustandsdiagramms, in dem jeder Zustand (i,j) gekennzeichnet ist durch

i = Anzahl der Anforderungen im System 1 , $i = 0,\ldots,S_1$
j = Anzahl der Anforderungen im System 2 , $j = 0,\ldots,S_2$

In diesem Diagramm sind die Ankünfte im System 1 (Ankunftsrate λ_1) durch Zustandsänderungen nach rechts, die Ankünfte direkt im System 2 (Ankunftsrate λ_2) durch Änderungen nach oben charakterisiert. Die Abgrenzungen des Zustandsraumes entstehen durch die endliche Speicherkapazität beider Systeme. Ein Bedienungsende im System 1 bewirkt eine Zustandsänderung nach links, falls die Anforderung mit Wahrscheinlichkeit 1-q das Verkehrsmodell verläßt, oder eine Zustandsänderung in der Diagonale, falls mit Wahrscheinlichkeit q ein Systemwechsel stattfindet.

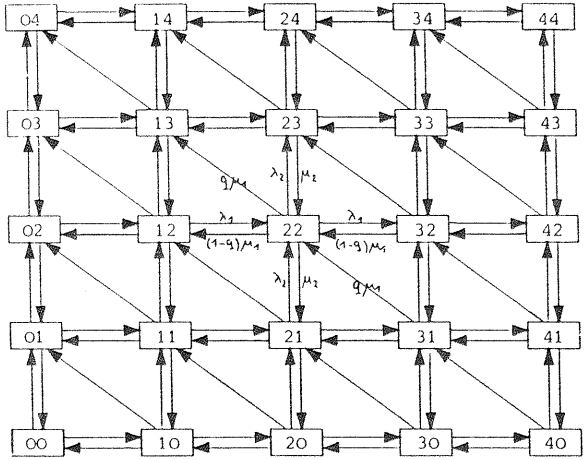


Bild 5.7: Zustandsdiagramm (Rückstau von Paketen)

Ein Bedienungsende im System 2 verursacht jeweils eine Zustandsänderung nach unten.

b) Zustandswahrscheinlichkeiten

Die stationären Zustandswahrscheinlichkeiten P_{ij} ermittelt man iterativ, die transienten Zustandswahrscheinlichkeiten $P_{ij}(t)$ mit dem Runge-Kutta-Verfahren. Die erforderlichen Gleichungen können entsprechend Abschnitt 4.2.3 aus dem Zustandsdiagramm entnommen werden.

Die Differentialgleichung für die Zustandswahrscheinlichkeit $P_{22}(t)$ lautet zum Beispiel:

$$\frac{d}{dt} P_{22}(t) = -[\lambda_1+\lambda_2(t)+\mu_1+\mu_2] \cdot P_{22}(t) + \lambda_1 \cdot P_{12}(t) + \lambda_2(t) \cdot P_{21}(t)$$
$$+ (1-q)\mu_1 \cdot P_{32}(t) + \mu_2 \cdot P_{23}(t) + q\mu_1 \cdot P_{31}(t) \quad . \qquad (5.7)$$

Die entsprechende Gleichung für den stationären Fall ergibt:

$$P_{22} = 1/(\lambda_1+\lambda_2+\mu_1+\mu_2) \cdot [\lambda_1 \cdot P_{12}+\lambda_2 \cdot P_{21}+(1-q)\mu_1 \cdot P_{32}+\mu_2 \cdot P_{23} + q\mu_1 \cdot P_{31}]. \quad (5.8)$$

c) Durchlaufprozeß

Für die Ermittlung der zeitabhängigen mittleren Durchlaufzeit $t_F(s)$ einer Testanforderung, die zum Zeitpunkt s im System 1 eintrifft und beide Systeme durchläuft, wird der Durchlaufprozeß benötigt. Dies hängt damit zusammen, daß ihre Gesamtdurchlaufzeit auch von späteren, direkt in System 2 eintreffenden Anforderungen beeinflußt wird.

Zur Berechnung dieser Durchlaufzeit wird folgende Zustandsdefinition für den Durchlaufprozeß eingeführt:

i = Anzahl der Vorgänger der Testanforderung im System 1,

j = { Anzahl aller Anforderungen im System 2, wenn die Testanforderung sich im System 1 befindet,
Anzahl der Vorgänger im System 2, wenn die Testanforderung sich im System 2 befindet.

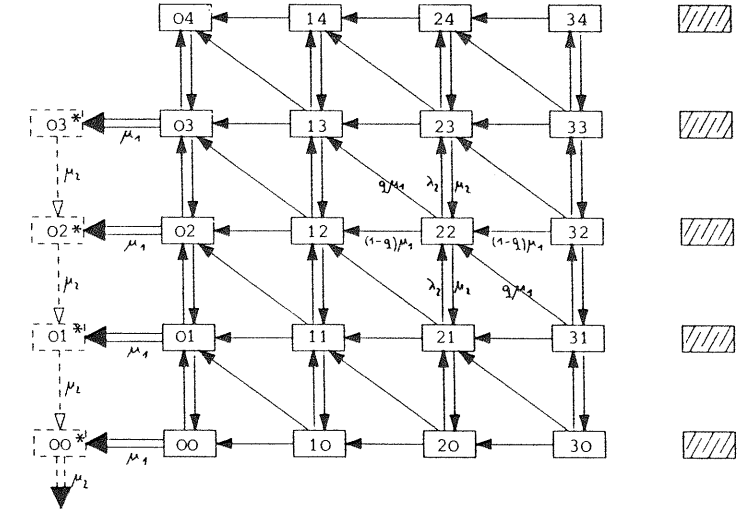


Bild 5.8: Durchlaufzustandsdiagramm (Rückstau von Paketen)

Bild 5.8 zeigt das mögliche Muster von Antreffzuständen mit sämtlichen Übergangsmöglichkeiten für den späteren Prozeßverlauf. Das Diagramm für den Durchlaufprozeß teilt sich auf in drei Teile:

- die schraffierten Systemzustände, die den Verlust der Testanforderung zufolge haben (volles System 1); diese Zustände gehören nach Abschnitt 4.2.4 zur Menge der absorbierenden Zustände H,
- die durchgezogenen Systemzustände für den Durchlaufprozeß im System 1; diese Zustände können direkt erreicht werden und gehören zur Menge $\hat{M}_1$,
- die gestrichelten Systemzustände für den Durchlaufprozeß im System 2; diese Zustände sind nur indirekt erreichbar und gehören somit zur Menge $\hat{M}_2$.

Der Durchlaufprozeß beginnt somit in einem durchgezogenen Zustand (i,j), der angibt, daß noch i Anforderungen vom System 1 zu bedienen sind bevor die Testanforderung ihre Bedienung erhält und daß sich zum Eintreffzeitpunkt s gerade j Anforderungen im System 2 befinden. Während des Aufenthaltes der Testanforderungen im System 1 kann sich der Zustand im System 2 noch beliebig ändern. Dies drückt sich im Zustandsdiagramm für den Durchlaufprozeß durch die vertikalen Übergänge aus. Das Vorrücken der Testanforderung im System 1 erfolgt durch einen Kolonnenwechsel in die linke Richtung, und zwar entweder horizontal mit Rate $(1-q)\mu_1$, falls ein Vorgänger das Verkehrsmodell verläßt, oder diagonal mit Rate $q\mu_1$, falls beide Systeme durchlaufen werden. In den Zuständen $(0,j)$, $j = 0,\ldots S_2$, wird die Testanforderung bedient, wobei im Zustand $(0,S_2)$ die Bedienung so oft wiederholt wird, bis die Testanforderung vom System 2 aufgenommen werden kann. Der Systemwechsel ist durch einen horizontalen dicken Übergangspfeil gekennzeichnet. Sobald die Testanforderung vom System 2 akzeptiert ist, fängt der Durchlaufprozeß im System 2 an: gestrichelte Zustände $(0,j)^*$, mit $j = 0,\ldots,S_2-1$. Ihre Durchlaufzeit kann nicht mehr von später eintreffenden Anforderungen beeinflußt werden. Der Durchlaufprozeß endet nach Bedienung der Testanforderung im System 2, d.h. nach Verlassen des Zustandes $(0,0)^*$.

Da die Durchlaufzeit der Testanforderung im System 2 aus einer Summe von unabhängigen Zufallsvariablen T_n mit negativ exponentieller Verteilung (Parameter μ_2) besteht, ist die Summenverteilung von k Bedienungszeiten durch eine Erlang-k-Verteilungsfunktion gegeben [Kleinrock (1975)]:

$$E_k(t) = P\{\sum_{n=1}^{k} T_n \le t\} = 1-e^{-\mu_2 t} \cdot \sum_{n=0}^{k-1} \frac{(\mu_2 t)^n}{n!} \quad . \tag{5.9}$$

Findet eine Testanforderung den Zustand $(0,j)^*$ vor, so müssen (j+1) negativ exponentiell verteile Bedienungsphasen durchlaufen werden, so daß die entsprechende bedingte komplementäre Durchlaufzeitverteilungsfunktion gegeben ist durch

$$E^c_{j+1}(t^*) = 1-E_{j+1}(t^*) = e^{-\mu_2 t^*} \cdot \sum_{n=0}^{j} \frac{(\mu_2 t^*)^n}{n!} \quad , \tag{5.10}$$

wobei $t^* = t-s$ das Zeitintervall zwischen Ankunftszeit s und Betrachtungszeitpunkt t ist.

Der gesamte Durchlaufprozeß wird durch ein System von Kolmogoroff-Rückwärts-Differentialgleichungen für die bedingten komplementären Durchlaufzeitfunktionen $f_{ij}(s,t)$ beschrieben. Die Differentialgleichungen lassen sich mit Hilfe der Regeln in Abschnitt 4.2.4 aus dem Durchlaufzustandsdiagramm aufstellen.

Damit entsteht das aus sechs Gleichungstypen bestehende Differentialgleichungssystem, das mit Hilfe der Runge-Kutta-Methode numerisch gelöst werden kann.

Für die bedingten komplementären Durchlaufzeitfunktionen der Durchlaufzustände in der untersten Zeile $(j = 0)$ gilt:

$$\frac{d}{ds} f_{00}(s,t) = -[\mu_1+\lambda_2(s)]\cdot f_{00}(s,t) + \lambda_2(s)\cdot f_{01}(s,t) + \mu_1 \cdot E^c_1(t^*) \tag{5.11a}$$

$$\frac{d}{ds} f_{i0}(s,t) = -[\mu_1+\lambda_2(s)]\cdot f_{i0}(s,t) + \lambda_2(s)\cdot f_{i1}(s,t) + (1-q)\mu_1 \cdot f_{i-1,0}(s,t)$$

$$+ q\mu_1 \cdot f_{i-1,1}(s,t) \qquad , i = 1,\ldots,S_1-1 \tag{5.11b}$$

Die Differentialgleichungen für die Zeilen $j = 1,\ldots,S_2-1$ lauten:

$$\frac{d}{ds} f_{Oj}(s,t) = -[\mu_1+\mu_2+\lambda_2(s)]\cdot f_{Oj}(s,t) + \lambda_2(s)\cdot f_{O,j+1}(s,t) + \mu_1\cdot E^C_{j+1}(t^*)$$

$$+\mu_2\cdot f_{O,j-1}(s,t) \qquad (5.12a)$$

$$\frac{d}{ds} f_{ij}(s,t) = -[\mu_1+\mu_2+\lambda_2(s)]\cdot f_{ij}(s,t) + \lambda_2(s)\cdot f_{i,j+1}(s,t) + (1-q)\mu_1\cdot f_{i-1,j}(s,t)$$

$$+ q\mu_1\cdot f_{i-1,j+1}(s,t) + \mu_2\cdot f_{i,j-1}(s,t) \qquad (5.12b)$$

$$, \; i = 1,\ldots,S_1-1 \quad .$$

Und schließlich gilt für die oberste Zeile ($j = S_2$):

$$\frac{d}{ds} f_{OS_2}(s,t) = -\mu_2\cdot f_{OS_2}(s,t) + \mu_2\cdot f_{O,S_2-1}(s,t) \qquad (5.13a)$$

$$\frac{d}{ds} f_{iS_2}(s,t) = -[(1-q)\mu_1+\mu_2]\cdot f_{iS_2}(s,t) + (1-q)\mu_1\cdot f_{i-1,S_2}(s,t)$$

$$+\mu_2\cdot f_{i,S_2-1}(s,t) \qquad , \; i = 1,\ldots S_1-1 \quad . \qquad (5.13b)$$

Die Gewichtung der bedingten komplementären Durchlaufzeitverteilungsfunktionen mit den entsprechenden Antreffwahrscheinlichkeiten gibt jetzt die gesuchte komplementäre Durchlaufzeitverteilungsfunktion für eine zum Zeitpunkt s eintreffende Testanforderung, die beide Warteschlangensysteme durchläuft:

$$F^C(s,t) = \sum_{i=0}^{S_1-1} \sum_{j=0}^{S_2} P_{ij}(s) \cdot f_{ij}(s,t) \qquad (5.14)$$

Interessiert man sich für die Durchlaufzeit bis die Testanforderung im System 2 eintrifft, so ist wegen der Blockierungsmöglichkeit ebenfalls eine Durchlaufprozeßbetrachtung erforderlich. Die entsprechenden Kolmogoroff-Rückwärts-Differentialgleichungen erhält man, wenn die Terme mit der Erlang-Funktion in Gl.(5.11a) bzw. Gl.(5.12a) weggelassen werden.

d) Charakteristische Verkehrsgrößen

Aus den Zustandswahrscheinlichkeiten können die folgenden charakteristischen Verkehrsgrößen ermittelt werden:

Verlustwahrscheinlichkeit im System 1 : (neue Anforderungen)

$$B_1(t) = \sum_{j=0}^{S_2} P_{S_1,j}(t) \qquad (5.15)$$

Verlustwahrscheinlichkeit im System 2 : (neue Anforderungen)

$$B_2(t) = \sum_{i=0}^{S_1} P_{i,S_2}(t) \qquad (5.16)$$

Blockierungswahrscheinlichkeit im System 2 : (Anforderungen von System 1 nach System 2)

$$B_{12}(t) = \frac{\sum_{i=1}^{S_1} P_{i,S_2}(t)}{\sum_{i=1}^{S_1} \sum_{j=0}^{S_2} P_{ij}(t)} \qquad (5.17)$$

Mittlere Belastung im System 1 :

$$E[X_1(t)] = \sum_{i=0}^{S_1} i \cdot \sum_{j=0}^{S_2} P_{ij}(t) \qquad (5.18)$$

Mittlere Belastung im System 2 :

$$E[X_2(t)] = \sum_{j=0}^{S_2} j \cdot \sum_{i=0}^{S_1} P_{ij}(t) \qquad (5.19)$$

Die mittlere Durchlaufzeit $t_F(s)$ einer Testanforderung, die zum Zeitpunkt s im System 1 eintrifft und beide Systeme durchläuft, erhält man durch numerische Integration der komplementären Durchlaufzeitverteilungsfunktion, gegeben in Gl.(5.14). Bezieht man das Ergebnis nur auf diejenigen Testanforderungen, die nicht vom System 1 abgewiesen werden, erhält man:

$$t_F(s \mid \text{akzeptiert}) = \frac{1}{1-B_1(s)} \cdot \int_{\tau=s}^{\infty} F^C(s,\tau) \cdot d\tau \qquad (5.20)$$

5.2.3 Numerische Ergebnisse

Als Beispiel seien die folgenden System- und Verkehrsparameter betrachtet:

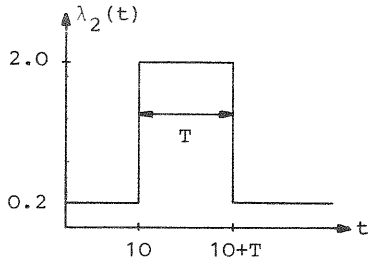
Systemkapazität : $S_1 = S_2 = 20$

Ankunftsrate 1 : $\lambda_1 = 0.7$

Ankunftsrate 2 : $\lambda_2(t)$

Bedienungsraten : $\mu_1 = \mu_2 = 1$

Verzweigungs-
wahrscheinlichkeit : $q = 5/7$



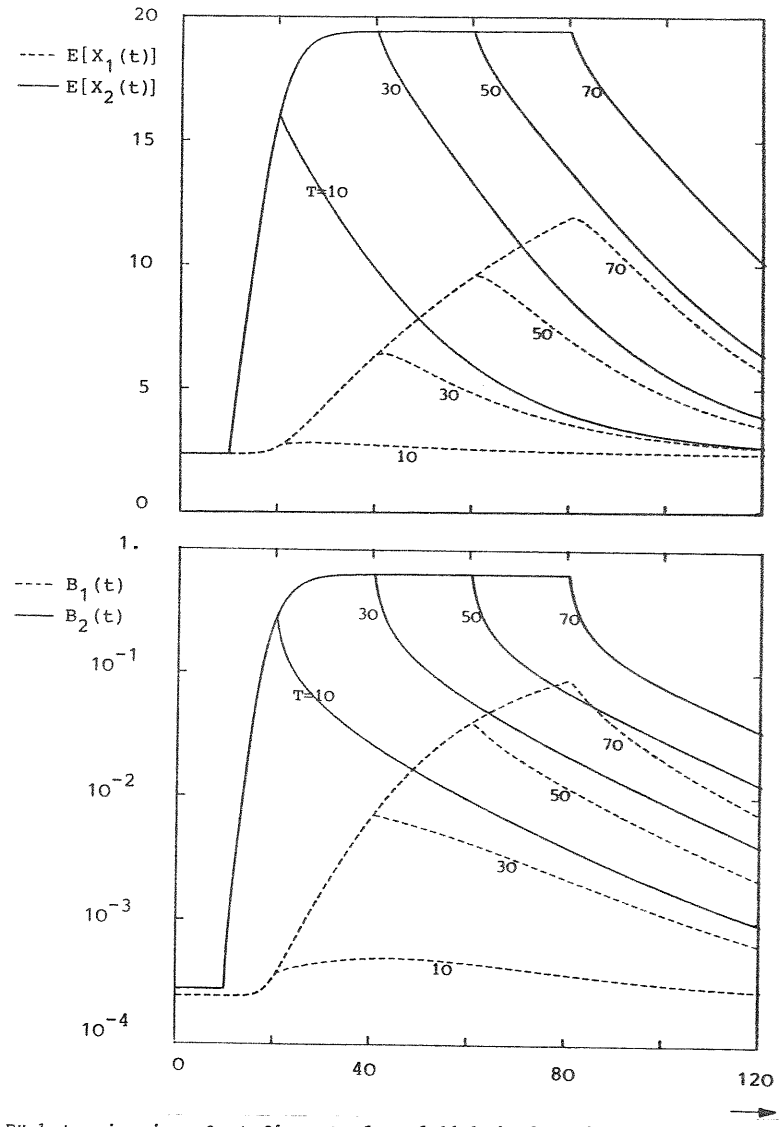
Betrachtet werden somit eine konstante Ankunftsrate λ_1 für System 1 und ein rechteckförmiger Überlastimpuls $\lambda_2(t)$ für System 2. Die Dauer der Überlast wird variiert (T = 10,30,50,70). Die Verzweigungswahrscheinlichkeit q ist so gewählt, daß im stationären Fall und unter Vernachlässigung von Verlusten der Verkehr in beiden Systemen identisch wäre.

a) Mittlere Systembelastung

Bild 5.9 gibt den zeitlichen Verlauf der mittleren Belastung in beiden Systemen wieder. Vor dem betrachteten Überlastimpuls sind die Systembelastungen zeichnerisch gleich. Durch die schlagartige Änderung der Ankunftsrate $\lambda_2(t)$ zum Zeitpunkt t = 10 nimmt die durchgezogene mittlere Systembelastung $E[X_2(t)]$ zunächst mit der Rate $(\lambda_{max}-\mu_2) + q\cdot\lambda_1 = (2-1) + 5/7\cdot(0.7) = 1.5$ zu. Bei einer Impulsdauer von T = 10 wird dadurch noch kein Sättigungszustand erreicht, so daß für diesen Fall die gestrichelte mittlere Systembelastung $E[X_1(t)]$ kaum beeinflußt wird. Dauert die Überlastsituation länger, so wird System 2 gesättigt und durch den Rückstau nimmt die mittlere Systembelastung $E[X_1(t)]$ zu.

b) Verlustwahrscheinlichkeit

In Bild 5.10 sind die zugehörigen Verlustwahrscheinlichkeiten der Anforderungsströme 1 und 2 dargestellt. Durch Verluste und Blockierung sind die stationären Ankunftsraten für beide Systeme etwas verschieden. Darüberhinaus ist durch die Blockierung der Ankunftsprozeß für System 2 kein Poisson-Prozeß mehr. Aus diesen Gründen sind die Anfangswerte von $B_1(t)$ und $B_2(t)$ nicht identisch. Im Ganzen zeigen diese Kurven jedoch ein ähnliches Verhalten wie die mittlere Systembelastungen.



Rückstau in einem 2-stufigen Tandemmodell bei einem im zweiten Warteschlangensystem auftretenden Überlastimpuls verschiedener Dauer T = 10,30,50,70.

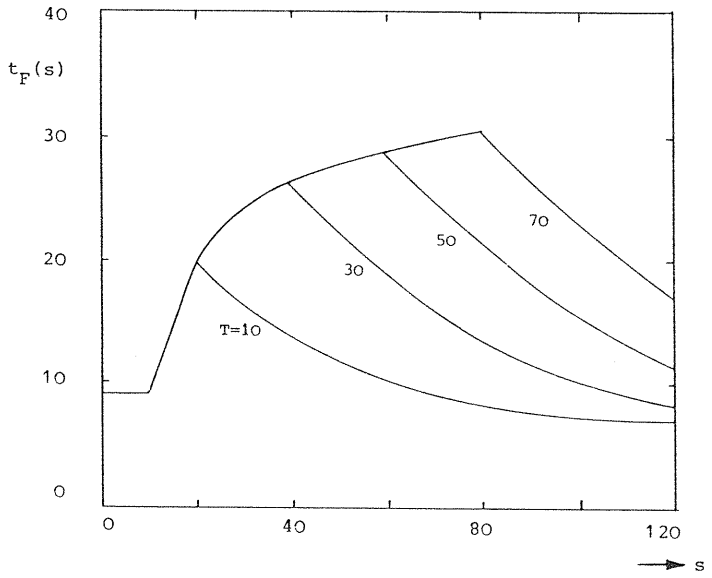
Bild 5.9: Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 5.10: Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.

Parameter: $S_1 = S_2 = 20$, $\lambda_1 = 0.7$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 2.0$, $q = 5/7$, $h_1 = h_2 = 1$.

c) Mittlere Durchlaufzeit

Bild 5.11 zeigt die mittlere Durchlaufzeit $t_F(s)$ einer Testanforderung, die zum Zeitpunkt s im System 1 eintrifft und beide Systeme durchläuft. Bedingt durch die schnelle Zunahme der mittleren Systembelastung $E[X_2(t)]$ im 2. System nimmt anfangs auch die mittlere Durchlaufzeit schnell zu. Danach wird der Kurvenverlauf im wesentlichen durch den aufbauenden Rückstau $E[X_1(t)]$ im 1. System bestimmt.



Rückstau in einem 2-stufigen Tandemmodell bei einem im zweiten Warteschlangensystem auftretenden Überlastimpuls verschiedener Dauer T = 10, 30, 50, 70.

Bild 5.11: Mittlere Durchlaufzeit durch beide Systeme $t_F(s)$ als Funktion vom Ankunftszeitpunkt s.
Parameter: $S_1 = S_2 = 20$, $\lambda_1 = 0.7$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 2.0$, $q = 5/7$, $h_1 = h_2 = 1$.

5.3 Die lawinenartige Überflutung des Netzes mit Paketkopien

Eine beachtenswerte Ursache für den betrieblichen Zusammenbruch von Paketvermittlungsnetzen liegt in der lawinenartigen Überflutung des Netzes mit Paketkopien verborgen, die nach Ablauf der Zeitbegrenzung (Time-Out) für die Paketquittierung automatisch erzeugt werden. Der hierfür verantwortliche Mechanismus der Mitkopplung wird in diesem Abschnitt mittels eines einfachen Verkehrsmodells vorgeführt. Eine ausführliche Untersuchung dieses Themas wird später in Kapitel 9 beschrieben.

5.3.1 Modellbeschreibung

Entsprechend Bild 5.12 betrachten wir ein Verkehrsmodell bestehend aus einem einstufigen Warteschlangensystem vom Typ M/M/1 mit einer von der Durchlaufzeit abhängigen Rückkopplungsschleife. Insbesondere wird ein Simulationsmodell betrachtet, so daß die stochastischen Fluktuationen voll zur Geltung kommen können.

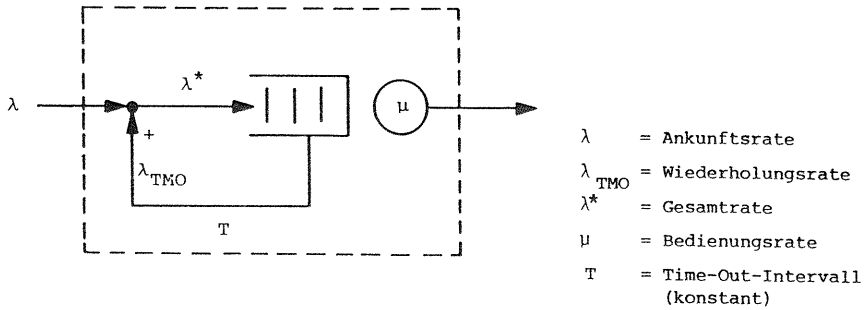


Bild 5.12: Verkehrsmodell: Lawinenartige Überflutung des Netzes mit Paketkopien.

Anforderungen treffen mit einer Ankunftsrate λ im System ein. Die Kapazität dieses Systems ist so groß gewählt, daß bis zum Zusammenbruch keine Verluste auftreten. Bei jeder Ankunft wird eine Zeitmarke für die maximal erlaubte Durchlaufzeit T gesetzt. Diese Zeitmarke wird beim Verlassen des Systems wieder gelöscht. Falls die Durchlaufzeit einer Anforderung die Zeitbegrenzung T überschreitet, wird dem System eine Kopie von dieser Anforderung übergeben und gleichzeitig wird die betreffende Zeitmarke

neu gesetzt. Dieser Vorgang kann sich so lange wiederholen, bis die Anforderung das System verlassen hat. Auf diese Weise kann jede Anforderung eine beträchtliche Blindlast erzeugen. Wird die Ankunftsrate der Kopien mit λ_{TMO} bezeichnet, dann beträgt die Gesamtankunftsrate $\lambda^* = \lambda + \lambda_{TMO}$.

5.3.2 Simulationsergebnisse

Bild 5.13 wurde während einer Simulation aufgezeichnet. Es gibt das Ablaufgeschehen kurz vor einem Systemzusammenbruch wieder. Die obere Aufzeichnung zeigt die momentane Anzahl der Anforderungen im System X(t), der untere Teil des Bildes gibt den darin enthaltenen Blindanteil $X_{TMO}(t)$ an. Während der Beobachtungszeit kann das Warteschlangensystem sich zweimal von einer Lastspitze erholen. Beim dritten Mal bricht es durch den Mitkopplungseffekt in kurzer Zeit zusammen.

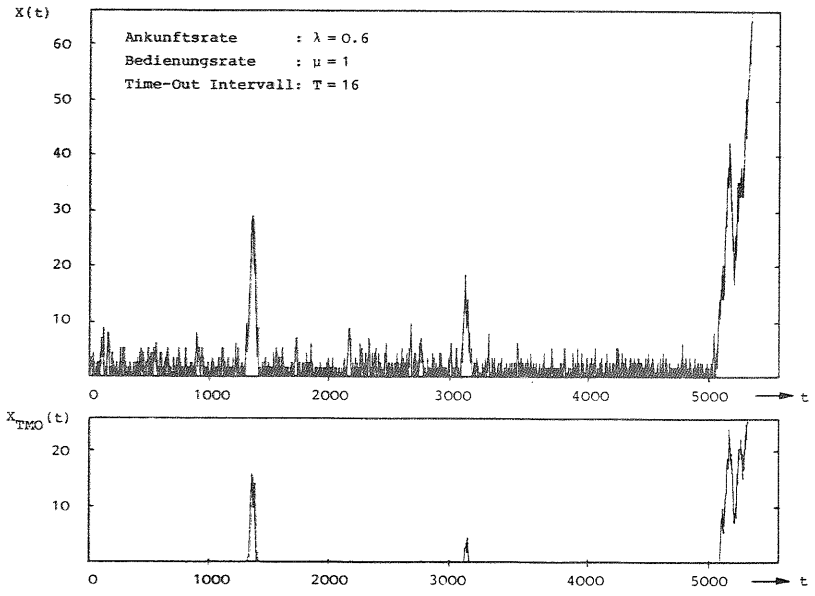


Bild 5.13: Simulationsaufzeichnung: Systemzusammenbruch.
X(t) : Anzahl der Anforderungen im System,
$X_{TMO}(t)$: Anzahl der Wiederholungen infolge eines Time-Outs.

5.4 Begrenzte Speicherkapazität und ihr Einfluß auf Verkehrsströme mit unsymmetrischen Verkehrsraten

Begrenzte Speicherkapaztäten beeinflussen in hohem Maße das Verkehrsverhalten von Paketvermittlungssystemen. Dieses kann besonders durch eine dynamische Betrachtungsweise veranschaulicht werden.

5.4.1 Modellbeschreibung

Das betrachtete Verkehrsmodell ist in Bild 5.14 abgebildet. Es handelt sich um ein einstufiges Warteschlangensystem mit einer Bedienungseinheit und zwei Klassen von Anforderungen, die sich eine begrenzte Speicherkapazität teilen müssen. Anforderungen der Klasse 1 - die weiter als 1-Anforderungen bezeichnet werden - treffen gemäß eines Poisson-Prozesses mit zeitabhängiger Ankunftsrate $\lambda_1(t)$ ein. Das Gleiche gilt für die 2. Klasse. Die Ankunftsrate der 2-Anforderungen sei mit $\lambda_2(t)$ bezeichnet.

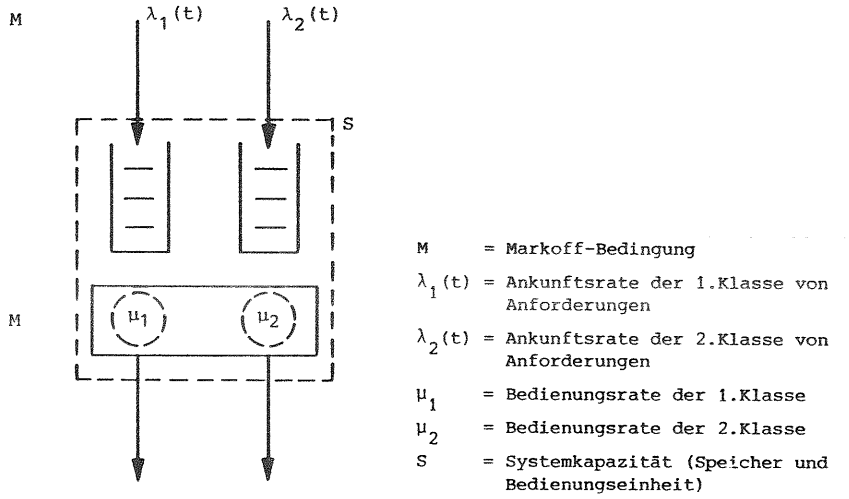


Bild 5.14: Verkehrsmodell: Einfluß einer begrenzten und gemeinsamen Systemkapazität bei zwei Klassen von Anforderungen.

Die Anforderungen werden von einer Bedienungseinheit mit einer klassenabhängigen Bedienrate μ_1 bzw. μ_2 behandelt. Die Bedienungszeiten seien negativ exponentiell verteilt. Die Anforderung in der Bedienungseinheit zählt zu der gemeinsamen Speicherkapazität S. Bezeichnet man die Anzahl der 1-Anforderungen im System mit X_1 und die Anzahl der 2-Anforderungen entsprechend mit X_2, so wird eine eintreffende Anforderung abgewiesen, wenn $X_1 + X_2 = S$.

Es werden im folgenden zwei Betriebsarten betrachtet: nichtunterbrechende Prioritäten und Abfertigung in zufälliger Reihenfolge. Bei nichtunterbrechenden Prioritäten werden 1-Anforderungen stets vor den 2-Anforderungen bedient, sie können jedoch eine 2-Anforderung in der Bedienungseinheit nicht unterbrechen. Bei der Abfertigung in zufälliger Reihenfolge wird die als nächste zu bedienende Anforderung zufällig aus allen wartenden Anforderungen ausgewählt.

5.4.2 Berechnungsverfahren

a) Zustandsdiagramm

Das Modell läßt sich mit Hilfe eines dreidimensionalen Markoff-Zustandsdiagramms lösen. Jeder Zustand sei gekennzeichnet durch (i,j,k):

i = Anzahl der 1-Anforderungen im System , i = 0,...,S
j = Anzahl der 2-Anforderungen im System , j = 0,...,S
k = Zustand der Bedienungseinheit
  k = 0 : leeres System
  k = 1 : Bedienung einer 1-Anforderung
  k = 2 : Bedienung einer 2-Anforderung

Bild 5.15 zeigt die Struktur dieses Zustandsdiagramms für ein Verkehrsmodell mit zwei Klassen von Anforderungen, die sich einen gemeinsamen Speicher teilen. In diesem Diagramm sind der leere Systemzustand und alle Zustände mit einer 1-Anforderung in der Bedienungseinheit durchgezogen dargestellt. Diejenigen Zustände, die eine Bedienung einer 2-Anforderung repräsentieren, sind hingegen gestrichelt gezeichnet. Das gleiche gilt für die zugehörigen Übergänge. Durch den gemeinsamen Speicher ist das Zustandsdiagramm in der Diagonale abgegrenzt.

Beispiel S = 5

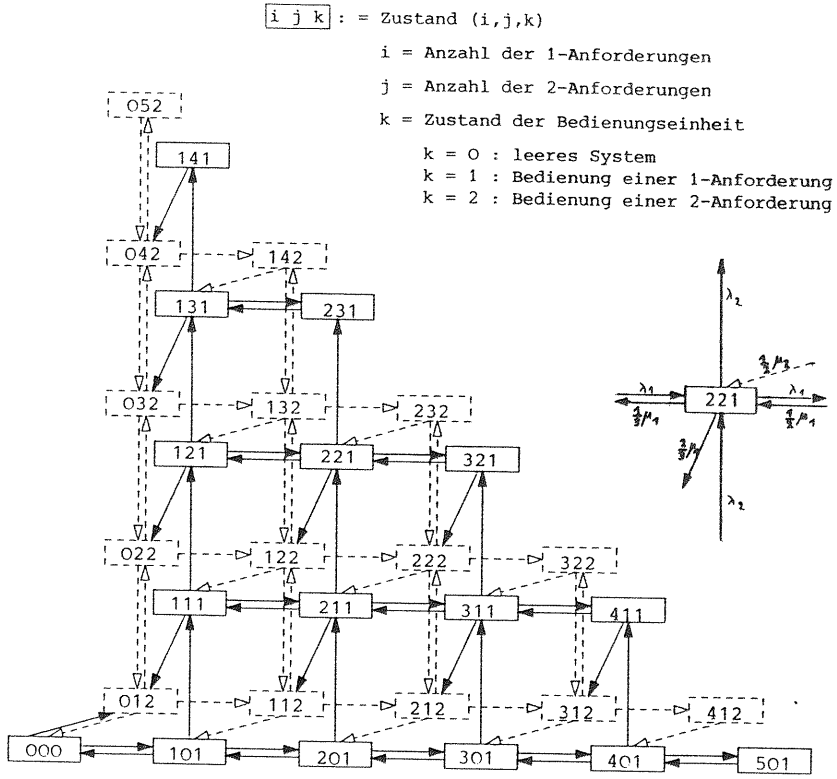


Bild 5.15: Zustandsdiagramm (Gemeinsame Systemkapazität und zwei Klassen von Anforderungen).

b) Abfertigungsstrategien

Durch entsprechende Wahl von den Klassenwechselwahrscheinlichkeiten q^m_{ijk} können verschiedene Abfertigungsstrategien berücksichtigt werden. Dabei ist

q^m_{ijk} die von i und j abhängige Wahrscheinlichkeit, daß nach Bedienung einer k-Anforderung (k = 1,2), eine Anforderung der Klasse m (m = 1,2) als nächste bedient wird oder das System in den leeren Zustand (m = 0) überwechselt.

Insbesondere gelten die folgenden Beziehungen

$$q^1_{ijk} + q^2_{ijk} = 1 \quad \text{und} \quad q^0_{101} = 1 \quad , \quad q^0_{012} = 1 \quad . \tag{5.21}$$

Wird die Abfertigung in zufälliger Reihenfolge betrachtet, so wechselt

- Zustand (i,j,1) mit i,j>0, beim Bedienungsende der 1-Anforderung mit Wahrscheinlichkeit $q^1_{ij1} = (i-1)/(i+j-1)$ über in den Zustand (i-1,j,1) und mit Wahrscheinlichkeit $q^2_{ij1} = j/(i+j-1)$ über in den Zustand (i-1,j,2).
- Zustand (i,j,2), mit i,j>0, beim Bedienungsende der 2-Anforderung mit Wahrscheinlichkeit $q^1_{ij2} = i/(i+j-1)$ über in den Zustand (i,j-1,1) und mit Wahrscheinlichkeit $q^2_{ij2} = (j-1)/(i+j-1)$ über in den Zustand (i,j-1,2).
- Zustand (1,0,1) bzw. (0,1,2) beim Bedienungsende der 1-Anforderung bzw. 2-Anforderung mit Wahrscheinlichkeit $q^0_{101} = 1$ bzw. $q^0_{012} = 1$ über in Zustand (0,0,0).

Beim Betriebsmodus nichtunterbrechende Prioritäten wechselt

- Zustand (i,j,1), mit i>1 und j>0, beim Bedienungsende der 1-Anforderung mit Wahrscheinlichkeit $q^1_{ij1} = 1$ über in den Zustand (i-1,j,1).
- Zustand (1,j,1), mit j>0, beim Bedienungsende der 1-Anforderung mit Wahrscheinlichkeit $q^2_{1j1} = 1$ über in den Zustand (0,j,2).
- Zustand (i,j,2), mit i,j>0, beim Bedienungsende der 2-Anforderung mit Wahrscheinlichkeit $q^1_{ij2} = 1$ über in den Zustand (i,j-1,1).
- Zustand (0,j,2), mit j>0, beim Bedienungsende der 2-Anforderung mit Wahrscheinlichkeit $q^2_{0j2} = 1$ über in den Zustand (0,j-1,2).
- Zustand (1,0,1) bzw. (0,1,2) beim Bedienungsende der 1-Anforderung bzw. 2-Anforderung mit Wahrscheinlichkeit $q^0_{101} = 1$ bzw. $q^0_{012} = 1$ über in Zustand (0,0,0).

c) Zustandswahrscheinlichkeiten

Wird die Zustandswahrscheinlichkeit des Zustandes (i,j,k) mit P_{ijk} bezeichnet, so können die entsprechenden Differentialgleichungen nach den im Abschnitt 4.2.3 zusammengefaßten Regeln aufgestellt werden. Beispielsweise lauten die Differentialgleichungen für die Zustände (2,2,1) und (1,3,2) wie folgt:

$$\frac{d}{dt} P_{221}(t) = -[\lambda_1(t) + \lambda_2(t) + \mu_1] \cdot P_{221}(t) + \lambda_1(t) \cdot P_{121}(t) + \lambda_2(t) \cdot P_{211}(t) + q^1_{321} \cdot \mu_1 \cdot P_{321}(t) + q^1_{232} \cdot \mu_2 \cdot P_{232}(t) \tag{5.22}$$

$$\frac{d}{dt} P_{132}(t) = -[\lambda_1(t) + \lambda_2(t) + \mu_2] \cdot P_{132}(t) + \lambda_1(t) \cdot P_{032}(t) + \lambda_2(t) \cdot P_{122}(t) + q^2_{231} \cdot \mu_1 \cdot P_{231}(t) + q^2_{142} \cdot \mu_2 \cdot P_{142}(t) \tag{5.23}$$

wobei für die Wahrscheinlichkeiten eines Klassenwechsels gilt:

nichtunterbrechende Prioritäten:

$$q^1_{321} = 1 \quad , \quad q^1_{232} = 1 \quad , \quad q^2_{231} = 0 \quad , \quad q^2_{142} = 0$$

zufällige Abfertigungsreihenfolge:

$$q^1_{321} = 1/2 \quad , \quad q^1_{232} = 1/2 \quad , \quad q^2_{231} = 3/4 \quad , \quad q^2_{142} = 3/4 \quad .$$

d) Charakteristische Verkehrsgrößen

Die interessierenden charakteristischen Verkehrsgrößen werden wie folgt aus den Zustandswahrscheinlichkeiten berechnet:

Verlustwahrscheinlichkeit $B(t) = B_1(t) = B_2(t)$

$$B(t) = \sum_{i+j=S} [P_{ij1}(t) + P_{ij2}(t)] \tag{5.24}$$

Mittlere Systembelastung durch 1-Anforderungen

$$E[X_1(t)] = \sum_i i \cdot \sum_j [P_{ij1}(t) + P_{ij2}(t)] \tag{5.25}$$

Mittlere Systembelastung durch 2-Anforderungen

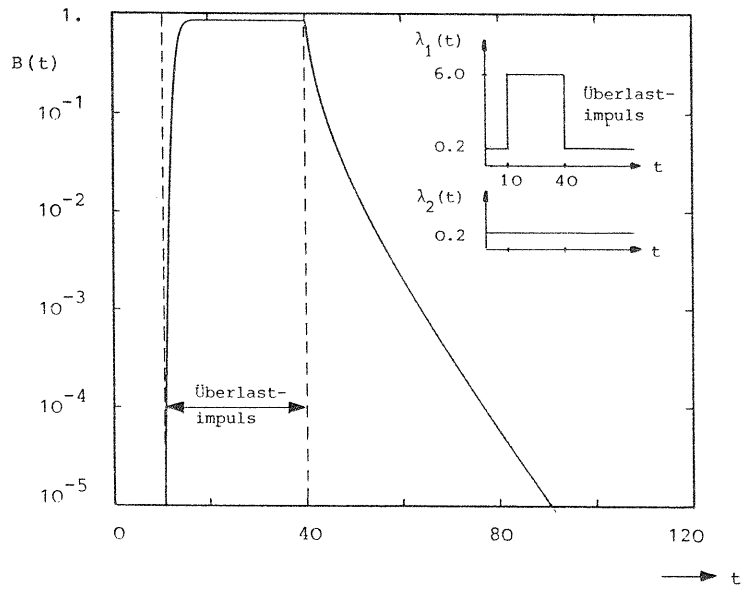
$$E[X_2(t)] = \sum_j j \cdot \sum_i [P_{ij1}(t) + P_{ij2}(t)] \tag{5.26}$$

Durchsatz von 1-Anforderungen $$D_1(t) = \mu_1 \cdot \sum_i \sum_j P_{ij1}(t) \qquad (5.27)$$

Durchsatz von 2-Anforderungen $$D_2(t) = \mu_2 \cdot \sum_i \sum_j P_{ij2}(t) \qquad (5.28)$$

5.4.3 Numerische Ergebnisse

Anhand eines numerischen Beispiels sei nun das typische Verkehrsverhalten von Systemen mit begrenzter Kapazität und unsymmetrischen Verkehrsströmen erläutert. Betrachtet wird ein System mit Kapazität S = 20. Die stationären Ankunftsraten beider Verkehrsströme seien identisch: $\lambda_1 = \lambda_2 = 0.2$. Ferner sollen beide Verkehrsklassen mit den gleichen Raten bedient werden: $\mu_1 = \mu_2 = 1$.



Gemeinsame und begrenzte Systemkapazität S mit zwei Anforderungsklassen, Überlastimpuls der 1. Anforderungsklasse (Impulsdauer T = 30).

Bild 5.16: Verlustwahrscheinlichkeit $B(t) = B_1(t) = B_2(t)$.
Parameter: $S = 20$, $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.2$, $h_1 = h_2 = 1$.

a) Verlustwahrscheinlichkeit

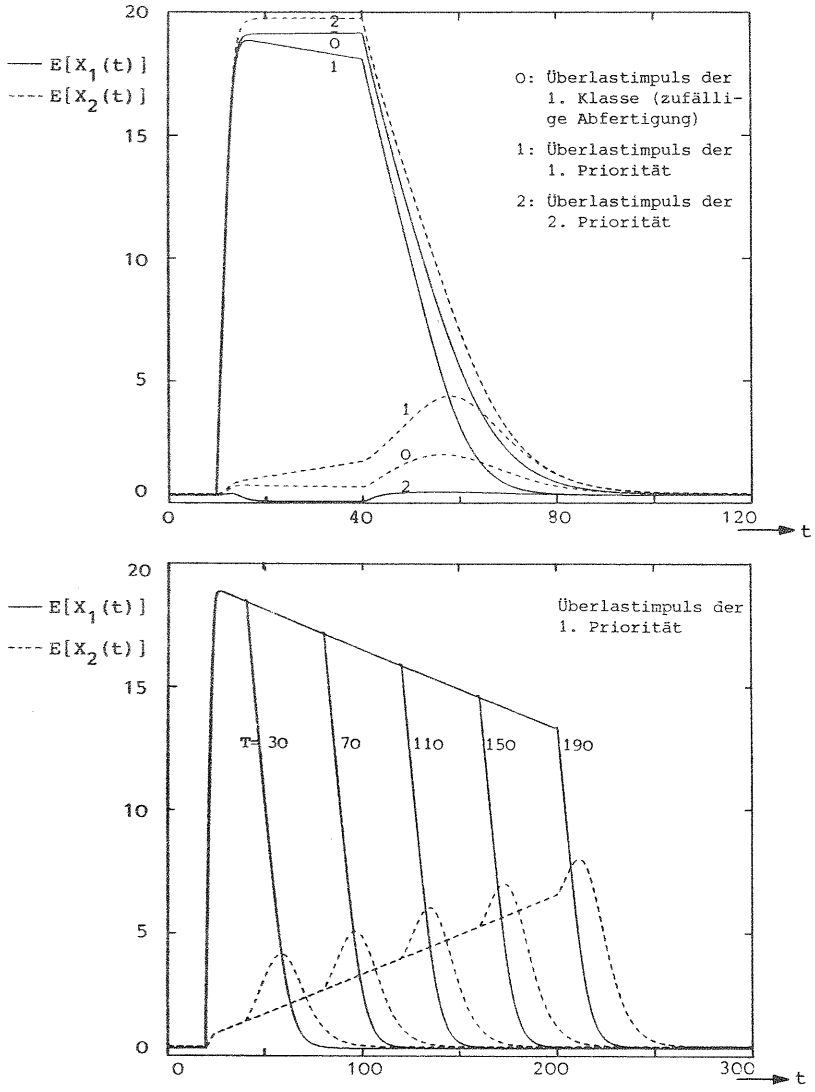
Bild 5.16 zeigt den Verlauf der Verlustwahrscheinlichkeit B(t), wenn das System einem sprungförmigen Überlastimpuls ausgesetzt wird, den eine der beiden Verkehrsklassen verursacht. Insbesondere wird $\lambda_1(t)$ nach einer rechteckförmigen Funktion verändert. Wegen des gemeinsamen Speichers ist die Verlustwahrscheinlichkeit für beide Verkehrsströme gleich. Sie ist unabhängig von der Abfertigungsreihenfolge und hängt lediglich von der Gesamtankunftsrate ab. Charakteristisch für das dynamische Systemverhalten ist die lange Nachwirkung eines Überlastimpulses.

b) Mittlere Systembelastung

In Bild 5.17 ist die mittlere Systembelastung für verschiedene Überlastsituationen vergleichend dargestellt.

Bei einer Abfertigung in zufälliger Reihenfolge (Kurven 0) wird im gesättigten Systemzustand die Kapazität des Systems in einem festen Verhältnis zwischen dem hohen (Klasse 1) und dem niedrigen (Klasse 2) Verkehrsstrom aufgeteilt. Dieses Verhältnis entspricht genau den Ankunftsraten; denn in dieser Proportion wird ein freiwerdender Speicherplatz neu belegt. Nach dem Überlastimpuls läßt sich eine leichte Erhöhung der mittleren Systembelastung $E[X_2(t)]$ feststellen, währenddessen die mittlere Systembelastung $E[X_1(t)]$ schnell abnimmt. Die kurze Erhöhung kommt dadurch zustande, daß während dieser Zeit weniger 2-Anforderungen abgewiesen werden, gleichzeitig aber die totale Systembelastung ihren normalen Wert noch nicht erreicht hat. Es bildet sich somit ein temporärer Rückstau.

Während eines Überlastimpulses der 2. Priorität (Kurven 2) sind kaum 1-Anforderungen in dem gemeinsamen Speicher vorhanden. Dies kann damit begründet werden, daß einerseits wegen ihrer hohen Ankunftsrate die 2-Anforderungen mit einer größeren Wahrscheinlichkeit einen freien Speicherplatz belegen, und andererseits aber die wenigen akzeptierten 1-Anforderungen bevorzugt behandelt werden. Auf die gleiche Weise wie vorher erhöht sich die Systembelastung des Verkehrsstromes mit der stationären Rate. Da es sich hierbei um die 1. Priorität handelt, ist diese Erhöhung jedoch sehr gering.



Gemeinsame und begrenzte Systemkapazität S mit zwei Anforderungsklassen (mit oder ohne nichtunterbrechende Prioritäten), Überlastimpuls einer Klasse. Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 5.17: Überlastimpuls bei unterschiedlichen Abfertigungsdisziplinen, Impulsdauer T = 30.

Bild 5.18: Überlastimpuls der 1. Prioritätsklasse (Impulsdauer T = 30, 70, 110, 150, 190). Parameter: S = 20, $\lambda_1(\infty) = \lambda_2(\infty) = 0.2$, $\lambda_{MAX} = 6.0$, $h_1 = h_2 = 1$.

Liegt ein Überlastimpuls der 1. Priorität vor (Kurven 1), so sollte festgehalten werden, daß ein Überschuß an 2-Anforderungen entsteht, der einer Aufnahme von neuen 1-Anforderungen entgegenwirkt. In gleichem Maße wie die mittlere Systembelastung $E[X_2(t)]$ ansteigt, nimmt die mittlere Systembelastung $E[X_1(t)]$ ab. Dies hängt damit zusammen, daß die 2-Anforderungen stets warten müssen bis sämtliche 1-Anforderungen bedient worden sind. Aus diesem Grund ist auch die Erhöhung der mittleren Systembelastung $E[X_2(t)]$ am Ende des Überlastimpulses größer als im Falle eines Betriebsmodus ohne Prioritäten.

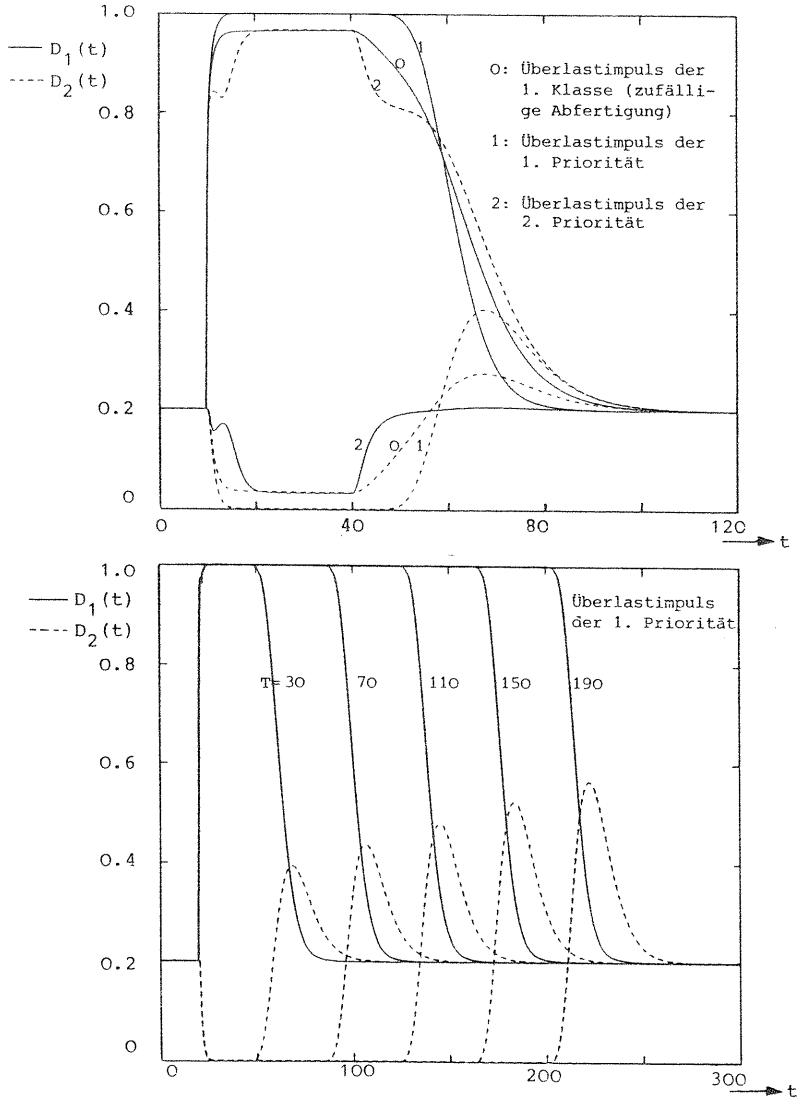
Bei einem lang andauernden Überlastimpuls der Priorität 1 wird der gemeinsame Speicher sogar völlig von den zurückgestellten 2-Anforderungen dominiert. Dieser Speicherbelegungsaustausch ist im Bild 5.18 für Überlastimpulse unterschiedlicher Zeitdauer T veranschaulicht. Die Geschwindigkeit, mit der die 2-Anforderungen sich rückstauen, hängt im wesentlichen von der Ankunftsrate λ_2 ab. Dadurch verringert sich der Speicheranteil der für 1-Anforderungen zur Verfügung steht. Im Grenzfall kann sogar eine 1-Anforderung nur aufgenommen werden, wenn gerade ein Speicherplatz frei wird. Um dieser Speicherplatzverstopfung durch 2-Anforderungen zuvorzukommen, sollten also einige Speicherplätze für 1-Anforderungen reserviert werden.

c) Durchsatz

Bild 5.19 gibt den Durchsatz beider Verkehrsklassen in Abhängigkeit der Zeit wieder.

Bei der Abfertigung in zufälliger Reihenfolge (Kurven 0) ist die Bedienungseinheit in einem festen, den Systembelastungen entsprechenden Verhältnis von beiden Verkehrsklassen belegt. Am Ende des Überlastimpulses erhöht sich der Durchsatz der benachteiligten Klasse 2 kurzzeitig über den stationären Wert hinaus (Abbau von zurückgestellten 2-Anforderungen).

Bei einem Überlastimpuls der 2. Priorität (Kurven 2) verhält sich der Durchsatz etwas komplexer. Zuerst wird der freie Speicherbereich sehr schnell mit 2-Anforderungen belegt, so daß der Durchsatz $D_2(t)$ ebenso schnell auf den zum $D_1(t)$ komplementären Wert 1.0-0.2 = 0.8 ansteigt. Da nicht immer



Gemeinsame und begrenzte Systemkapazität S mit zwei Anforderungsklassen (mit oder ohne nichtunterbrechende Prioritäten), Überlastimpuls einer Klasse. Durchsatz $D_1(t)$ bzw. $D_2(t)$.

Bild 5.19: Überlastimpuls bei unterschiedlichen Abfertigungsdisziplinen, Impulsdauer T = 30.

Bild 5.20: Überlastimpuls der 1. Prioritätsklasse (Impulsdauer T = 30,70, 110,150,190). Parameter: S = 20, $\lambda_1(\infty) = \lambda_2(\infty) = 0.2$, $\lambda_{MAX} = 6.0$, $h_1 = h_2 = 1$.

1-Anforderungen anwesend sind, kann Durchsatz $D_2(t)$ sogar noch weiter zunehmen. Ein Bedienungszyklus von 1-Anforderungen wird nun öfters durch eine Bedienung einer 2-Anforderung verzögert. Nach dieser Verzögerung werden sämtliche zwischenzeitlich angekommenen 1-Anforderungen bedient, so daß $D_2(t)$ für einen kurzen Moment wieder abnimmt. Danach wird ein freiwerdender Speicherplatz mit höherer Wahrscheinlichkeit von einer 2-Anforderung belegt, so daß dadurch der Durchsatz $D_2(t)$ bis auf einen maximalen Wert anwächst. Dieser Wert hängt vom Verhältnis der Ankunftsraten ab. Sobald der Überlastdruck der 2-Anforderungen verschwunden ist, fällt Durchsatz $D_2(t)$ zuerst rapide bis zum Wert 0.8 ab. Grund dafür ist die rasch abnehmende Verlustwahrscheinlichkeit B(t), so daß weniger Anforderungen abgewiesen werden und der Durchsatz $D_1(t)$, komplementär zu $D_2(t)$ wieder zunehmen kann. Wenn $D_1(t)$ seinen Endwert 0.2 erreicht, nimmt $D_2(t)$ erneut schnell ab. Da während der Systemerholungsphase die mittlere Systembelastung $E[X_1(t)]$ etwas höher ist, liegt der Durchsatz $D_1(t)$ in diesem Zeitintervall knapp über dem stationären Wert.

Bei einem Überlastimpuls der 1. Priorität (Kurven 1) sind sämtliche 2-Anforderungen beinahe völlig blockiert und dementsprechend ist ihr Durchsatz äußerst klein, währenddessen der Durchsatz $D_1(t)$ sehr groß ist. Dieses Durchsatzverhältnis bleibt auch eine gewisse Zeit nach Beendigung der Überlast bestehen. In dieser Zeit werden alle 1-Anforderungen, die noch im Speicher vorhanden sind, bedient. Danach verursachen die zurückgestellten 2-Anforderungen eine kurze und intensive Erhöhung des Durchsatzes $D_2(t)$. Diese Verzögerungszeit ist umso kürzer und die Erhöhung im Durchsatz umso heftiger, je länger der Überlastimpuls der 1-Anforderungen angedauert hat. Denn mit zunehmender Überlastdauer sind weniger 1-Anforderungen vorhanden, dafür sind aber um so mehr 2-Anforderungen im System zurückgestellt (vgl. auch Bild 5.18). Dies ist abschließend in Bild 5.20 veranschaulicht.

6. ÜBERLASTABWEHR DURCH REGELUNG DER NETZZUGÄNGE

Als erste Maßnahme zur Überlastabwehr wird die Regelung der Netzzugänge diskutiert. Die dazu erforderlichen Informationen erhält man entweder aus Meldungen von anderen Netzknoten oder aus eigenen Zustandsinformationen. Im ersten Fall handelt es sich um eine globale, im zweiten um eine lokale Überlastabwehrstrategie.

6.1 Allgemeines

In diesem Kapitel wird der Netzzugangsverkehr aufgrund einer lokalen Überlastabwehrstrategie geregelt. Als Entscheidungsgrundlage zur Einleitung einer bremsenden Maßnahme dient die momentane Speicherbelegung im betrachteten Netzendknoten. Dabei werden die Pakete eingeteilt in Pakete die bereits vom Netz übermittelt wurden (Transitpakete), und in Pakete, die direkt aus dem Anschlußnetz kommen (Ursprungspakete). Die in der Literatur vorgeschlagenen Strategien unterscheiden sich hauptsächlich durch die Form der Paketabweisung und die Anwendung der Prioritätsregeln. Ziel dieser lokalen Überlastabwehrstrategien ist es, das Entstehen von Blindlast durch Paketabweisungen im Netzinneren zu unterbinden. Dazu ist es notwendig, den Transitpaketen einen größeren Speicherbereich einzuräumen, so daß, wenn Überlastsituationen in einem Netzendknoten entstehen, bereits durch Speicherzuteilungsstrategien die weitere Zufuhr von Paketen aus dem Anschlußnetz gebremst oder gestoppt werden kann. Obwohl die Überlastabwehrstrategie nur auf lokalen Zustandsinformationen basiert, werden durch den Rückstau-Effekt indirekt auch Überlastsituationen in anderen Netzteilen berücksichtigt.

Die Limitierung von Ursprungspaketen aufgrund der Speicherbelegung im Netzendknoten wird im Zusammenhang mit dem Pufferklassen-Konzept simulativ untersucht in [Raubold/Haenle (1976), Giessler/Haenle/König/Pade (1978)]. Die Simulationen zeigen, daß für eine vorgegebene Netztopologie und Verkehrsmatrix ein optimaler Wert für das Abweisen von Ursprungspaketen (IBL, Input Buffer Limit) existiert. Sowohl höhere als auch niedrigere Werte verursachen eine beträchtliche Verringerung im Durchsatz.

In [Lam/Reiser (1979)] wird mit Hilfe eines Warteschlangennetzmodells die IBL-Strategie analytisch untersucht. Eine entsprechende simulative Untersuchung für ein Paketvermittlungsnetz erfolgt in [Lam/Lien (1981)]. Aus beiden Arbeiten geht hervor, daß der Richtwert für das Verhältnis β zwischen Abweisgrenze S_2 und Gesamtspeicher S wie folgt gewählt werden muß, damit ein maximaler Netzdurchsatz erreicht werden kann:

$$\beta = \frac{S_2}{S} < \frac{\alpha}{\bar{H}} \quad . \tag{6.1}$$

Hierbei ist $\alpha<1$ ein Skalierungsfaktor, um Verkehrsunsymmetrien auszugleichen und $\bar{H}$(Hops) die mittlere Zahl von Netzknoten die durchlaufen werden.

In [Schwartz/Saad (1979)] werden die Ursprungspakete dann abgewiesen, wenn die Gesamtbelegung im Speicher (Ursprungspakete plus Transitpakete) die Grenze S_2 überschreitet. Da den Transitpaketen zusätzlich $S-S_2$ Speicherplätze zur Verfügung stehen, wird die Strategie als Free Buffer Allocation oder als Additional Buffer Allocation (ABA) bezeichnet. Die Untersuchung erfolgt analytisch mit Hilfe eines einstufigen Warteschlangenmodells. Eine ähnliche Strategie wird für einen Netzknoten in [Kamoun (1981)] als Warteschlangennetz untersucht. In [Saad/Schwartz (1980)] wird die IBL-Strategie mit nichtunterbrechenden Prioritäten (IBL-PRIO) betrachtet und mit den Strategien IBL und ABA bezüglich Durchsatz, Durchlaufzeit und Leistung (d.h. Durchsatz/Durchlaufzeit) verglichen. Der Vergleich zeigt, daß die Strategie ABA das beste Überlastabwehrverhalten aufweist. Danach folgen IBL-PRIO und IBL.

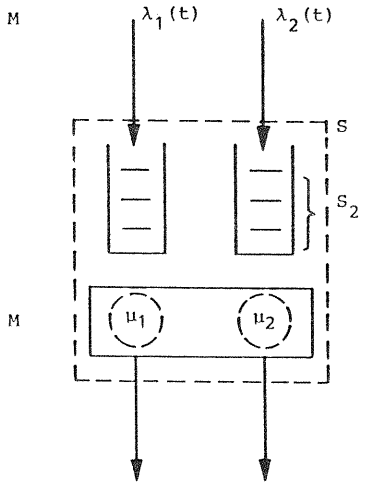
Im folgenden werden die Strategien IBL und ABA, mit oder ohne nichtunterbrechenden Prioritäten, aufgrund anderer Kriterien einander gegenübergestellt. Einerseits ist dies die Verlustwahrscheinlichkeit als Maß für die Anzahl der Paketwiederholungen, andererseits die dynamische Reaktion auf eine Lastspitze.

6.2 Modellbeschreibung

Um die verschiedenen Strategien miteinander vergleichen zu können, wird das in Bild 6.1 abgebildete Verkehrsmodell betrachtet. Es besteht aus einem einstufigen Warteschlangensystem mit zwei Klassen von Anforderungen, die von einer einzigen Bedienungseinheit mit unterschiedlichen Bedienungsraten abgefertigt werden. Den Anforderungen steht eine gemeinsame Speicherkapazität S, die auch die bediente Anforderung enthält, zur Verfügung. Ferner existiert eine Speicherbelegungsgrenze S_2 für die 2. Anforderungsklasse. Damit unterliegen die Verkehrsströme folgenden Speicherbelegungseinschränkungen:

S : verfügbare Speicherkapazität für die Anforderungen der 1. Klasse (1-Anforderungen),

S_2: verfügbare Speicherkapazität für die Anforderungen der 2. Klasse (2-Anforderungen).



M = Markoff-Bedingung

$\lambda_1(t)$ = Ankunftsrate der 1.Klasse (Transit Pakete)

$\lambda_2(t)$ = Ankunftsrate der 2.Klasse (Ursprungspakete)

μ_1 = Bedienungsrate der 1.Klasse

μ_2 = Bedienungsrate der 2.Klasse

S = Systemkapazität (Speicher und Bedienungseinheit)

S_2 = Belegungsgrenze für die 2.Klasse

Bild 6.1: Verkehrsmodell: Überlastabwehr durch Regelung der Netzzugänge.

Das Verkehrsmodell stellt die gesamte Speicherkapazität in einem Netzendknoten dar, wobei die Transitpakete (1-Anforderungen) bzw. die Ursprungspakete (2-Anforderungen) den Speicher mit der Rate μ_1 bzw. μ_2 verlassen. Die Bedienung der Anforderungen der beiden Klassen geschieht entweder nach nichtunterbrechenden Prioritäten oder nach einer Abfertigung in zufälliger Reihenfolge. Mit der letzten Abfertigungsstrategie erfolgt die Bedienung der beiden Klassen proportional zu ihrem momentanen Belegungszustand. Die beiden Verkehrsquellen sind als zeitabhängige Poisson-Prozesse mit Ankunftsrate $\lambda_1(t)$ bzw. $\lambda_2(t)$ modelliert. Die zufälligen Bedienungszeiten sind als negativ exponentiell verteilt angenommen. Für die mittlere Bedienungszeit gilt $h_1 = 1/\mu_1$ bzw. $h_2 = 1/\mu_2$.

Wird die Anzahl der 1-Anforderungen bzw. der 2-Anforderungen im System mit der Zufallsvariable X_1 bzw. X_2 beschrieben, so wird die Aufnahme von Anforderungen durch folgende Regeln bestimmt:

- 1-Anforderungen werden akzeptiert, wenn

$$X_1 + X_2 < S \tag{6.2}$$

- 2-Anforderungen werden akzeptiert, wenn

$$\text{IBL} : X_1 + X_2 < S \quad \text{und} \quad X_2 < S_2 \tag{6.3}$$

$$\text{ABA} : X_1 + X_2 < S_2 \quad .$$

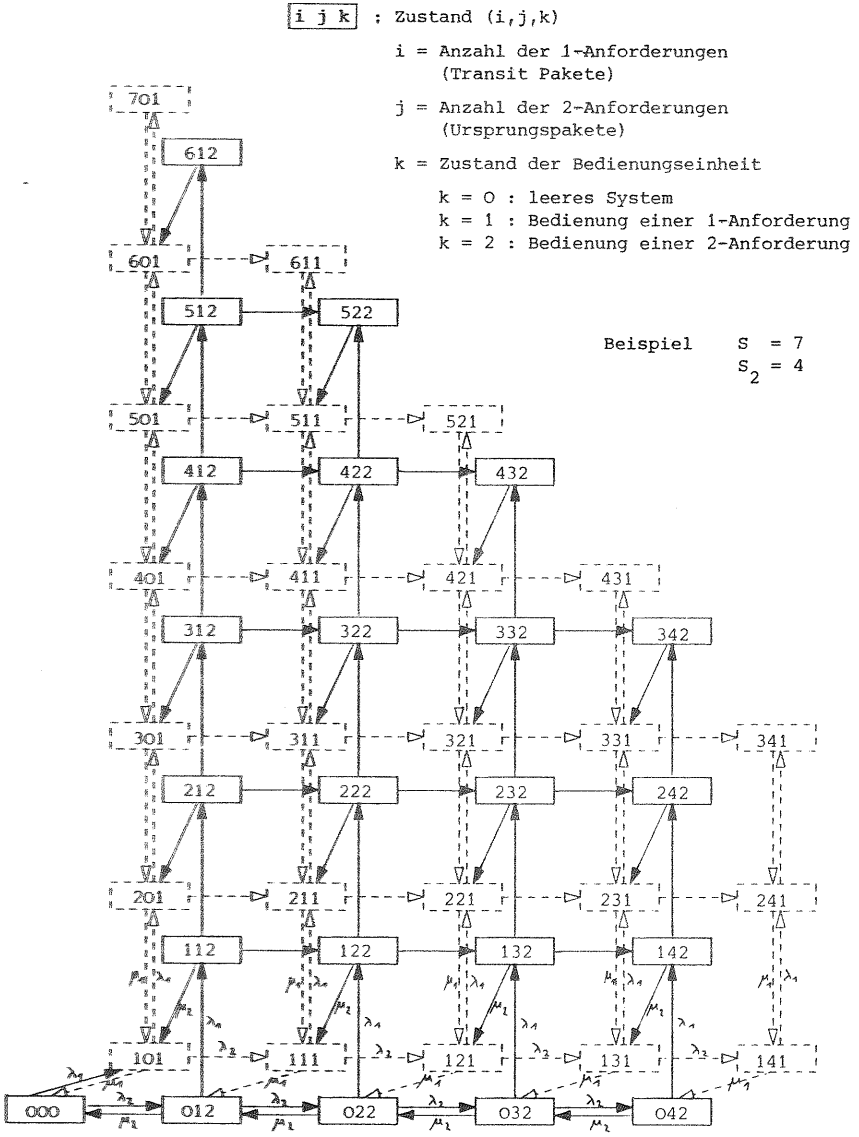


Bild 6.2: Zustandsdiagramm (Strategie IBL-PRIO: Input Buffer Limit mit Prioritäten).

6.3 Modellanalyse

a) Zustandsdiagramm

Als Beispiel für die betrachteten Betriebsarten dieses Verkehrsmodells ist in Bild 6.2 das dreidimensionale Zustandsdiagramm für die Strategie IBL-PRIO dargestellt. Die Zustände sind charakterisiert durch das Zahlentripel (i,j,k) mit

i = Anzahl der 1-Anforderungen im System , $i = 0,\ldots,S$
j = Anzahl der 2-Anforderungen im System , $j = 0,\ldots,S_2$
k = Zustand der Bedienungseinheit

k = 0 : leeres System
k = 1 : Bedienung einer 1-Anforderung
k = 2 : Bedienung einer 2-Anforderung

Der leere Systemzustand, die Zustände mit einer 2-Anforderung in der Bedienungseinheit und die zugehörigen Übergangspfeile sind durchgezogen gezeichnet. Die Zustände und Übergangspfeile, die zu einer Bedienung einer 1-Anforderung gehören, sind gestrichelt dargestellt. Aufgrund der Speicherbegrenzung $S_2 < S$ für die 2-Anforderung weist das Zustandsdiagramm eine entsprechende Begrenzung durch die Zustände $(i+1,S_2,1)$ und $(i,S_2,2)$, mit $i = 0,\ldots,S-S_2$, auf. Die momentane Belegungsmöglichkeit der 1-Anforderungen, die den ganzen Speicher belegen dürfen, hängt von der Anzahl der im System vorhandenen 2-Anforderungen ab. Dies drückt sich durch die diagonale Abgrenzung des Zustandsdiagramms aus. Das Diagramm zeigt ferner die Zustandsübergänge, die im Betriebsmodus für nichtunterbrechende Prioritäten vorkommen können. Beim Bedienungsende einer 2-Anforderung wird stets, falls anwesend, eine 1-Anforderung als nächste bedient (Wechsel von der durchgezogenen zur gestrichelten Zustandsebene). Warten jedoch keine 1-Anforderungen auf Bedienung, dann wird eine weitere 2-Anforderung bedient (unterste durchgezogene Zustandszeile bzw. Zustände $(0,j,2)$, $j = 1,\ldots,S_2$) oder das System geht in den leeren Zustand über. Ist eine 1-Anforderung in der Bedienungseinheit, so finden, falls noch weitere 1-Anforderungen anwesend sind, alle Übergänge in der gestrichelten Ebene statt. Ein Wechsel von der gestrichelten zur durchgezogenen

Zustandsebene erfolgt nur, wenn die letzte im System vorhandene 1-Anforderung bedient worden ist (unterste gestrichelte Zustandszeile bzw. Zustände $(1,j,1)$, $j = 0,\ldots,S_2$).

Die Zustandsdiagramme für die anderen Überlastabwehrstrategien erhält man durch Hinzufügen (zufällige Bedienungsreihenfolge, vgl. Abschnitt 5.4) bzw. Weglassen (Strategie ABA) von Zustandsübergängen. Im letzteren Fall sind die Zustände $(i,S_2,1)$, mit $i = 1,\ldots,S-S_2$, in der rechten gestrichelten Zustandskolonne nicht erreichbar.

b) Zustandswahrscheinlichkeiten

Anhand des Zustandsdiagramms (Bild 6.2) können die Gleichungen für die stationären Zustandswahrscheinlichkeiten P_{ijk} sowie die Differentialgleichungen für die transienten Zustandswahrscheinlichkeiten $P_{ijk}(t)$ aufgestellt werden. Die Regeln dazu sind im Abschnitt 4.2.3 zusammengestellt. Im stationären Fall werden sie mit dem iterativen Verfahren, im transienten Fall mit dem Runge-Kutta-Verfahren bestimmt (Abschnitte 4.4.1 bzw. 4.4.3).

Als Beispiel wird der Zustand (1,0,1) betrachtet, für den die folgende Differentialgleichung gilt:

$$\frac{d}{dt} P_{101}(t) = -\,[\lambda_1(t)+\lambda_2(t)+\mu_1]\cdot P_{101}(t) + \lambda_1(t)\cdot P_{000}(t) + \mu_1 \cdot P_{201}(t) + \mu_2 \cdot P_{112}(t) \quad . \tag{6.4}$$

Und für den stationären Fall gilt:

$$P_{101} = 1/(\lambda_1+\lambda_2+\mu_1)\cdot[\lambda_1\cdot P_{000} + \mu_1\cdot P_{201} + \mu_2\cdot P_{112}] \quad . \tag{6.5}$$

c) Charakteristische Verkehrsgrößen

Aus den Zustandswahrscheinlichkeiten lassen sich die im nächsten Abschnitt betrachteten charakteristischen Verkehrsgrößen ableiten:

- Verlustwahrscheinlichkeit für 1-Anforderungen:

$$B_1(t) = \sum_{i+j=S} [P_{ij1}(t) + P_{ij2}(t)] \tag{6.6}$$

- Verlustwahrscheinlichkeit für 2-Anforderungen:

Strategie IBL: (6.7a)

$$B_2(t) = \sum_{i+j=S} [P_{ij1}(t) + P_{ij2}(t)] + \sum_{i=1}^{S-S_2-1} P_{i,S_2,1}(t) + \sum_{i=0}^{S-S_2-1} P_{i,S_2,2}(t)$$

Strategie ABA:

$$B_2(t) = \sum_{i+j\geq S_2} [P_{ij1}(t) + P_{ij2}(t)] \tag{6.7b}$$

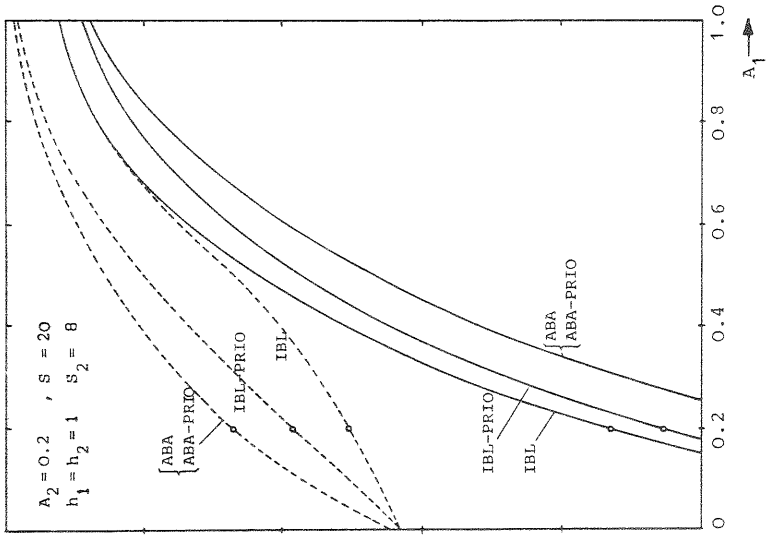
- Mittlere Systembelastung und Durchsatz wie in Abschnitt 5.4.2.d.

6.4 Numerische Ergebnisse

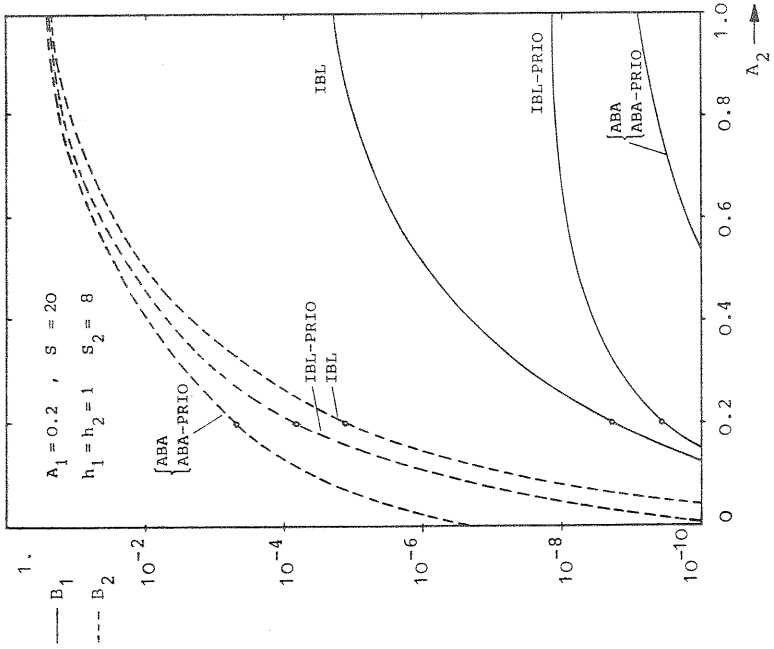
Als Ergänzung zu den Resultaten, die in der angegebenen Literatur zu finden sind, wird in diesem Abschnitt die Verlustwahrscheinlichkeit als Maß für die Anzahl Wiederholungen und die dynamischen Eigenschaften der Strategien IBL, IBL-PRIO, ABA und ABA-PRIO diskutiert.

6.4.1 Stationäre Ergebnisse

Betrachtet sei ein Warteschlangensystem mit Speicherkapazität $S = 20$ und einer Speicherbelegungsgrenze $S_2 = 8$ zur Limitierung der 2-Anforderungen (Ursprungspakete). Nach Gl.(6.1) durchlaufen die Pakete dabei im Mittel $\bar{H} = 2.5$ Netzknoten. Ferner seien die mittleren Bedienungszeiten $h_1 = h_2 = 1$.



Regelung der Netzzugänge.
Bild 6.4: Verlustwahrscheinlichkeit B_1 bzw. B_2 als Funktion vom Verkehrsangebot A_1.



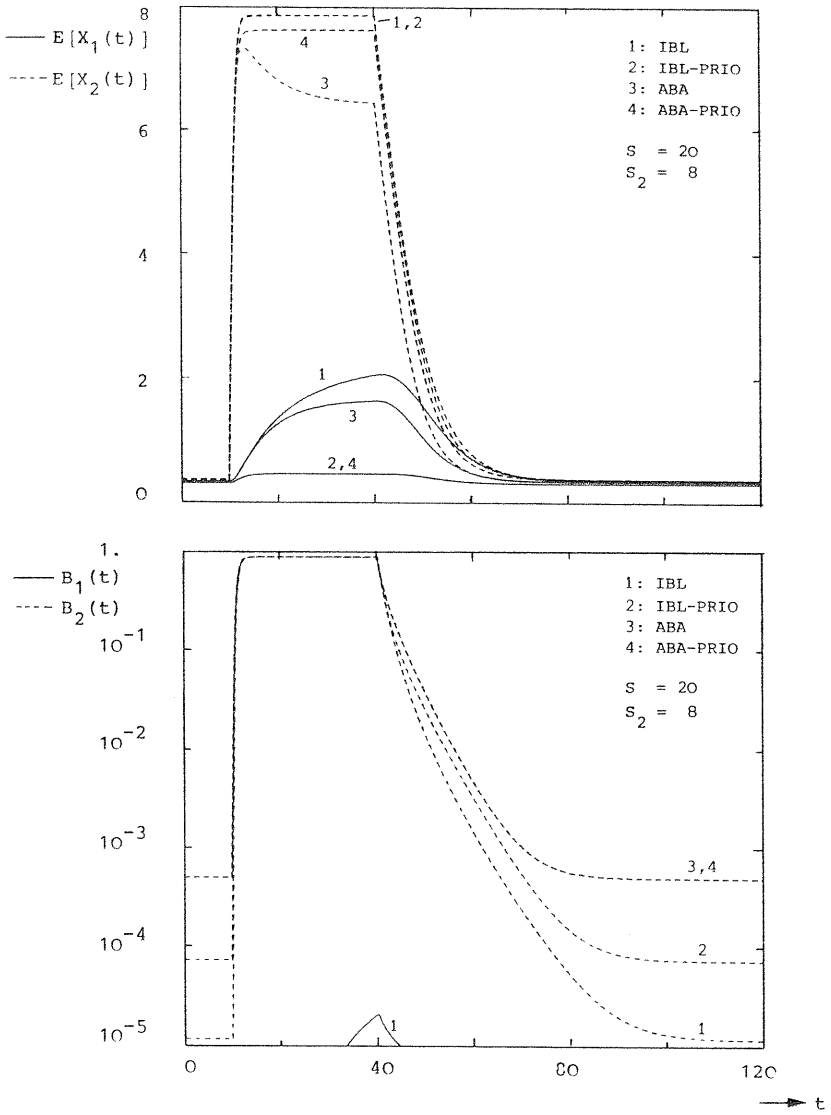
Regelung der Netzzugänge.
Bild 6.3: Verlustwahrscheinlichkeit B_1 bzw. B_2 als Funktion vom Verkehrsangebot A_2.

a) Verlustwahrscheinlichkeit in Abhängigkeit vom Verkehrsangebot A_2

In Bild 6.3 sind die Kurven für die Verlustwahrscheinlichkeiten B_1 und B_2 der verschiedenen Strategien als Funktion vom Verkehrsangebot A_2 aufgetragen. Als Betriebspunkt seien die Verkehrsangebote $A_1 = 0.2$ und $A_2 = 0.2$ für die 1-Anforderungen bzw. 2-Anforderungen betrachtet. Die entsprechenden Verlustwahrscheinlichkeiten sind speziell markiert. Wie man erkennt, liegt für diese Verkehrsverhältnisse die Verlustwahrscheinlichkeit B_1 der Strategien IBL und IBL-PRIO in der Größenordnung von 10^{-9}, die der Strategien ABA und ABA-PRIO bei etwa 10^{-11}. Bedingt durch die gewählten Systemparameter sind die Werte der beiden letzten Kurven gleich. Die Verlustwahrscheinlichkeit B_2 liegt in der Größenordnung von 10^{-4}. Wird das Verkehrsangebot A_2 erhöht, so nimmt B_2 bei allen Strategien schnell zu, währenddessen die Verlustwahrscheinlichkeit B_1 auch bei einem hohen Verkehrsangebot noch kleine Werte aufweist. Die Zunahme von B_1 ist bei der Strategie IBL am größten. Deshalb lassen die Kurven die Schlußfolgerung zu, daß aufgrund der geringen Verluste für 1-Anforderungen sogar bei einem hohen Angebotsdruck von 2-Anforderungen (Ursprungspakete) eine Betriebsweise mit einer niedrigen Wiederholungsrate für die 1-Anforderungen (Transitpakete) gewährleistet ist.

b) Verlustwahrscheinlichkeit in Abhängigkeit von Verkehrsangebot A_1

Bild 6.4 zeigt die entsprechenden Kurven, wenn das Verkehrsangebot der 1. Anforderungsklasse variiert wird. Auch hier ist der Betriebspunkt mit $A_1 = 0.2$ und $A_2 = 0.2$ markiert. Wird das Verkehrsangebot A_1 erhöht, so läßt sich feststellen, daß sowohl die Verlustwahrscheinlichkeit B_1 als auch die Verlustwahrscheinlichkeit B_2 schnell ansteigen. Darüberhinaus nähern sich die Kurven B_1 und B_2 der Strategien IBL bei wachsendem Angebot A_1. Für die Strategie IBL sind die Verluste bei einem Verkehrsangebot $A_1 > 0.7$ für beide Anforderungsklassen gleich. Für IBL-PRIO gilt dies für $A_1 > 1.6$. Dagegen bleibt bei den Strategien ABA bzw. ABA-PRIO immer eine Differenz bestehen (siehe auch Bild 6.9). Die Kurven lassen somit erkennen, daß bei einer von



Regelung der Netzzugänge: Überlastimpuls der 2. Anforderungsklasse (Ursprungspakete), Impulsdauer T = 30.

Bild 6.5: Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 6.6: Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.
Verkehrsparameter: $\lambda_1 = 0.2$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

1-Anforderungen (Transitpakete) verursachten Überlastung die weitere Annahme von 2-Anforderungen (Ursprungspakete) entsprechend verringert wird. Bei der Strategie ABA ist die Schutzwirkung am besten. Da die 1-Anforderungen den gesamten Speicher monopolisieren und somit die Abweisungen bewirken, können Wiederholungen bei den 1-Anforderungen auf diese Weise nicht vermieden werden.

6.4.2 Transiente Ergebnisse

Zur Charakterisierung der dynamischen Eigenschaften werden die gleichen Systemparameter betrachtet. Die stationären Ankunftsraten seien $\lambda_1 = \lambda_2 = 0.2$ und die mittleren Bedienungszeiten $h_1 = h_2 = 1$. Ferner wird jeweils die Ankunftsrate einer der beiden Klassen gemäß einer Rechteckfunktion mit Höchstwert $\lambda_{max} = 6.0$ geändert. Der Überlastimpuls beginnt bei der normierten Zeit $t = 10$ und endet bei $t = 40$.

6.4.2.1 Überlastimpuls der 2. Anforderungsklasse (Ursprungspakete)

Als erstes wird die Systemreaktion auf einen Überlastimpuls der 2. Anforderungsklasse untersucht.

a) Mittlere Systembelastung

In Bild 6.5 ist der zeitliche Verlauf der mittleren Systembelastung für die beiden Anforderungsklassen aufgezeichnet. Bedingt durch den Überlastimpuls erreicht die mittlere Systembelastung $E[X_2(t)]$ schnell ihren Maximalwert, der durch die Schranke $S_2 = 8$ bestimmt wird. Bei der Strategie IBL und IBL-PRIO kann dieser Bereich völlig ausgenutzt werden. Bei den Strategien ABA und ABA-PRIO wird das Aufnahmevermögen von 2-Anforderungen durch anwesende 1-Anforderungen reduziert und entsprechend beeinflußt die mittlere Systembelastung $E[X_1(t)]$ den Verlauf von $E[X_2(t)]$. Im Bild ist im Überlastbereich der komplementäre Kurvenverlauf deutlich erkennbar. Die mittlere Systembelastung $E[X_1(t)]$ wird bei den prioritätsorientierten Strategien lediglich aufgrund der größeren Restbedienungswahrscheinlichkeit einer 2-Anforderung erhöht. Für die beiden Strategien ohne Prioritäten nimmt $E[X_1(t)]$ zuerst linear zu und wird dann etwas

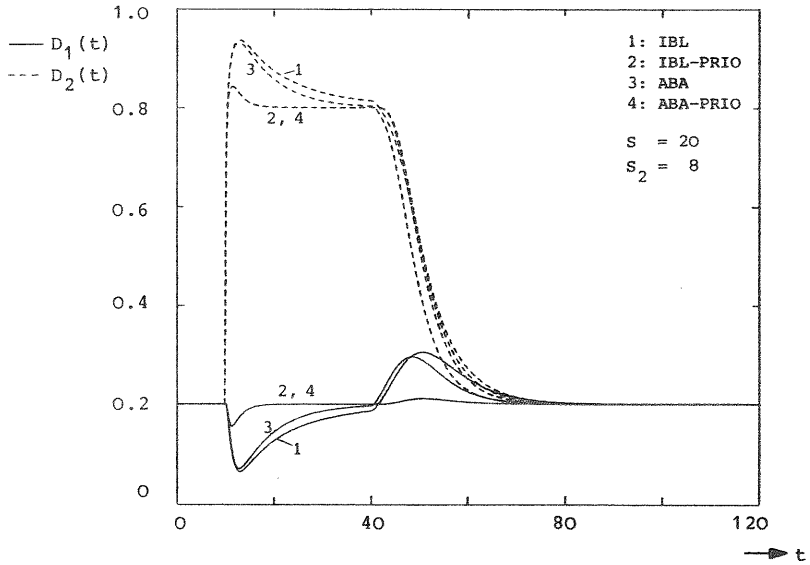
abgeflacht. Dieser abflachende Kurvenverlauf ergibt sich aus der Abfertigungsstrategie, bei der aus allen wartenden Anforderungen die nächste Anforderung zufällig ausgewählt wird. Dadurch wird diejenige Anforderungsklasse mit der längsten Warteschlange schneller bedient. Die Kurve für die Strategie ABA flacht hierbei etwas schneller ab, weil die 1-Anforderungen mit weniger 2-Anforderungen konkurrieren müssen.

b) Verlustwahrscheinlichkeit

Bild 6.6 zeigt das entsprechende Verhalten der Verlustwahrscheinlichkeiten $B_1(t)$ und $B_2(t)$. Wie auch bereits aus Bild 6.3 für stationäre Verhältnisse hervorging, wird die Verlustwahrscheinlichkeit $B_1(t)$ nur geringfügig von der Überlast der 2-Anforderungen beeinflußt. Im vorliegenden Bild ist sie für die Strategie IBL noch gerade erkennbar.

c) Durchsatz

In Bild 6.7 ist der Durchsatz der beiden Anforderungsklassen dargestellt. In Bezug auf den Kurvenverlauf kann zwischen den Strategien mit und denjenigen ohne Prioritäten unterschieden werden. Für die Strategien IBL-PRIO und ABA-PRIO steigt der Durchsatz $D_2(t)$ fast direkt bis über den zu $D_1(t)$ komplementären Wert an. Nach dieser kurzen Erhöhung bleibt $D_2(t)$ während des Überlastimpulses auf dem Wert 0.8 und er sinkt nach Verschwinden der Überlast wieder schnell auf den ursprünglichen Wert. Entsprechend verhält sich $D_1(t)$: zuerst eine kurze Durchsatzminderung, dann während der Überlast auf den ursprünglichen Wert und anschließend noch eine geringfügige Durchsatzsteigerung. Der Durchsatz $D_2(t)$ für die Strategien IBL und ABA fängt bei einem hohen Wert an und sinkt dann allmählich ab. Bei der Strategie ABA erfolgt diese Abnahme sogar noch etwas schneller. Dieser Verlauf hängt wieder mit der Abfertigungsstrategie zusammen, in der die Anforderungsklasse mit der längsten Warteschlange am schnellsten bedient wird. Der entsprechende Durchsatz $D_1(t)$ verläuft während der Überlast komplementär und ist kurz nach dem Überlastimpuls kurzfristig wegen der Bedienung der rückgestauten 1-Anforderungen etwas höher.



Regelung der Netzzugänge: Überlastimpuls der 2. Anforderungsklasse (Ursprungspakete), Impulsdauer T = 30.

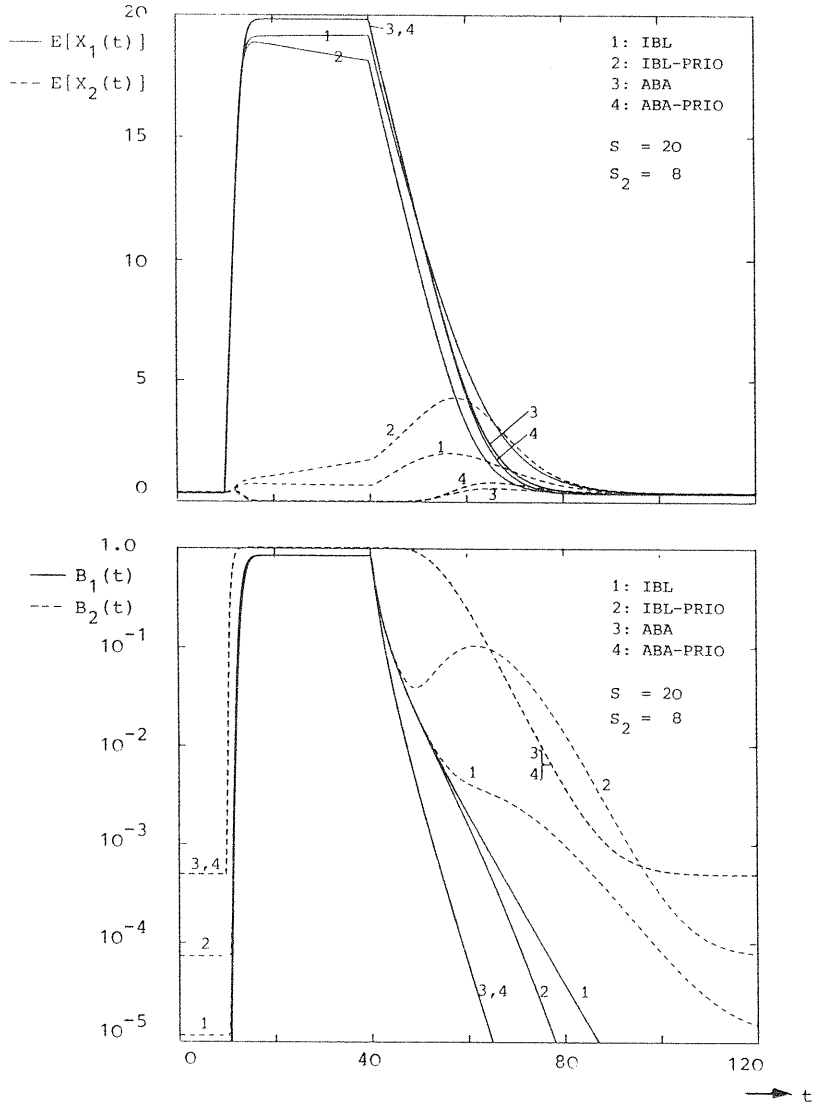
Bild 6.7: Durchsatz $D_1(t)$ bzw. $D_2(t)$.
Verkehrsparameter: $\lambda_1 = 0.2$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

6.4.2.2 Überlastimpuls der 1. Anforderungsklasse (Transitpakete)

Die nächsten Ergebnisse gelten für einen Überlastimpuls der 1. Anforderungsklasse.

a) Mittlere Systembelastung

In Bild 6.8 ist die mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$ veranschaulicht. Da die 1-Anforderungen den gesamten Speicher belegen können, liegen die gleichen Verhältnisse wie bei einem gemeinsamen Speicher vor. Die Kurven der Strategien IBL bzw. IBL-PRIO sind deshalb auch identisch mit den Kurven 0 bzw. 1 in Bild 5.17. Sie werden dort ausführlich diskutiert. Anders ist es bei den Strategien ABA und ABA-PRIO. Be-



Regelung der Netzzugänge: Überlastimpuls der 1. Anforderungsklasse (Transitpakete), Impulsdauer T = 30.

Bild 6.8: Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 6.9: Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.2$, $h_1 = h_2 = 1$.

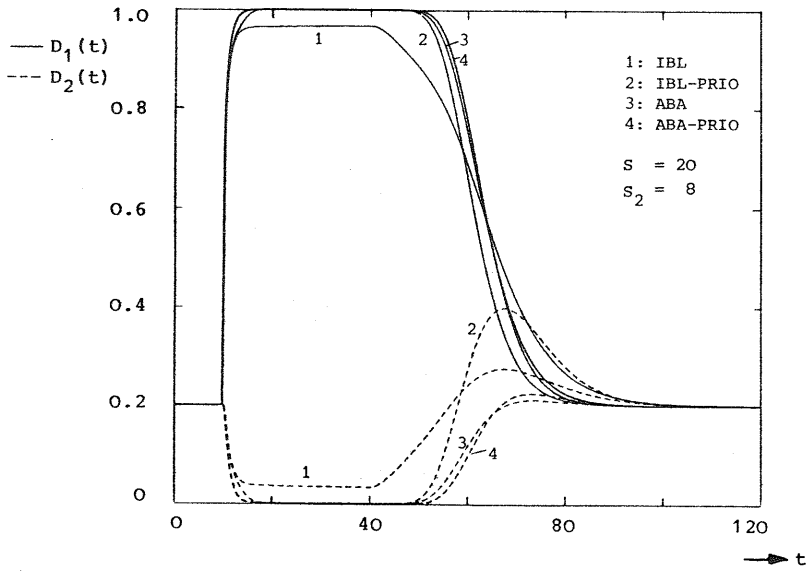
dingt durch die strengeren Abweismaßnahmen für die 2-Anforderungen wird während der Überlastsituation der ganze Speicher von 1-Anforderungen belegt. Die mittlere Systembelastung $E[X_2(t)]$ wird dadurch auf etwa Null reduziert, und zwar solange bis $E[X_1(t)]$ auf etwa die Speicherbelegungsgrenze $S_2 = 8$ abgefallen ist. Ab diesem Punkt fällt auch die mittlere Systembelastung $E[X_1(t)]$ bei der Strategie ABA-PRIO etwas schneller ab als bei der Strategie ABA. Entsprechend zeigt $E[X_2(t)]$ in dieser Zeitspanne bei ABA-PRIO einen etwas höheren Rückstau als bei ABA.

b) Verlustwahrscheinlichkeit

In Bild 6.9 wird der zeitliche Verlauf der beiden Verlustwahrscheinlichkeiten diskutiert. Als Referenzkurve dient die Verlustwahrscheinlichkeit $B_1(t) = B_2(t)$ für den gemeinsamen Speicher in Bild 5.15. Diese Referenzkurve deckt sich mit der Kurve für die Strategie IBL solange $B_1(t) = B_2(t)$ ist. Danach liegt sie geringfügig rechts von der durchgezogenen Verlustwahrscheinlichkeit $B_1(t)$ der betrachteten Strategie. Das Auseinanderlaufen der Verlustkurven $B_1(t)$ und $B_2(t)$ bei der Strategie IBL und bereits etwas früher bei IBL-PRIO hängt mit der kurzzeitigen Überhöhung der mittleren Systembelastung $E[X_2(t)]$ nach Verschwinden des Überlastimpulses zusammen (vgl. Bild 6.8). Durch diesen Rückstau stehen weniger freie Speicherplätze für die 2-Anforderungen zur Verfügung, so daß die Wahrscheinlichkeit für eine Abweisung größer wird. Bei der Strategie IBL-PRIO weist die Kurve sogar ein ausgeprägtes zweites Maximum auf. Dagegen ist bei der Strategie IBL lediglich eine Verzögerung für den Kurvenabfall zu erkennen. Entsprechend schneller kann dadurch die Verlustwahrscheinlichkeit $B_1(t)$ für diese beiden Strategien abfallen. Die Verlustwahrscheinlichkeit $B_2(t)$ für die Strategie ABA bzw. ABA-PRIO läßt auch hier die gute Überlastabwehreigenschaft durch sofortige totale Abweisung von 2-Anforderungen gut erkennen. Ihre Wirkung setzt sich über eine geraume Zeit nach dem Überlastimpuls noch fort. Durch diese länger andauernde Schutzwirkung kann sich der Überschuß an 1-Anforderungen schnell abbauen, so daß die Verlustwahrscheinlichkeit $B_1(t)$ auch sehr schnell ihren stationären Wert erreicht.

c) Durchsatz

Bild 6.10 zeigt die Durchsatzkurven. Der Verlauf für die Strategie IBL bzw. IBL-PRIO entspricht dem der Kurve 0 bzw. 1 im Bild 5.19. Ihre Diskussion erfolgt in Abschnitt 5.4.3.c. Es werden deshalb hier die Kurven der Strategien ABA und ABA-PRIO betrachtet. Gemäß den vorherigen Erläuterungen in Bezug auf mittlere Systembelastung und Verlustwahrscheinlichkeit beansprucht der Durchsatz $D_1(t)$ beinahe sofort die gesamte Bedienungszeit und dieser maximale Durchsatz bleibt auch nach Ende der Überlastsituation einige Zeit beibehalten. Da sich während der Systemerholungsphase nur eine geringe Belastung von 2-Anforderungen aufbauen kann, fehlt bei der Durchsatzkurve $D_2(t)$ die stoßartige Durchsatzerhöhung.



Regelung der Netzzugänge: Überlastimpuls der 1. Anforderungsklasse (Transitpakete), Impulsdauer T = 30.

Bild 6.10: Durchsatz $D_1(t)$ bzw. $D_2(t)$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.2$, $h_1 = h_2 = 1$.

6.5 Schlußfolgerung

Literatur und die durchgeführten dynamischen Untersuchungen zeigen, daß durch Regelung der Netzzugänge aufgrund der momentanen Speicherbelegung im betreffenden Netzendknoten eine wirksame und leicht implementierbare Überlastabwehrstrategie möglich ist. Insbesondere wird bei der Strategie ABA (Additional Buffer Allocation) der Netzzugang sofort gesperrt, wenn eine Überlastspitze durch die netzinternen Pakete (Transitpakete) verursacht wird.

7. ÜBERLASTABWEHR DURCH AUSLAGERUNG

In Kapitel 6 wurden Überlastabwehrstrategien betrachtet, die den Paketverkehr direkt an der Netzübergabestelle regeln können. Die abgewiesenen Pakete werden zwar später wiederholt, aber die Auswirkung dieses zusätzlichen Verkehrsangebots beschränkt sich auf eine Verringerung der effektiven Paketverarbeitungskapazität der betreffenden Netzendknoten. Weitere Betriebsmittel des Paketvermittlungsnetzes sind nicht von Paketwiederholungen betroffen. Dies ist jedoch nicht mehr der Fall, wenn Pakete im Netz selbst durch einen Speicherüberlauf in einem Netzknoten oder durch eine Überlastabwehrmaßnahme [Kamoun (1981)] abgewiesen werden. In diesem Fall wird je nach Netzimplementierung ein Paketwiederholungsvorgang im Nachbarknoten, im Netzendknoten oder beim Benutzer selbst eingeleitet. Unabhängig von der Implementierung jedoch verringert sich die Effizienz der Netzbetriebsmittel, und zwar gehen Übertragungs-, Verarbeitungs- und Speicherkapazitäten verloren. Dies kann zur Verschärfung der Überlastsituation führen. Damit die vom Netz bereits akzeptierten Pakete mit möglichst geringen Kapazitätsverlust zum Ziel befördert werden, benötigt man schnell ansprechende Überlastabwehrstrategien.

7.1. Die Foreground-Background Strategie

In diesem Kapitel wird eine Überlastabwehrstrategie, basierend auf einer selektiven Auslagerung von Paketen, vorgestellt und untersucht. Sie wird im weiteren als Foreground-Background (FG-BG) Strategie bezeichnet. Im wesentlichen soll diese Strategie bewirken, daß die im Netz unvermeidlich auftretenden, kurzen Überlastspitzen von den Netzknoten selbst aufgefangen werden können. Auf diese Weise soll vermieden werden, daß die Überlastsituation eine unkontrollierte Auswirkung auf Nachbarknoten hat. Der erzielte Gewinn an Netzkapazität ergibt sich aus der verbesserten Annahmestrategie für die Pakete, so daß weniger Paketwiederholungen erforderlich sind.

Als Voraussetzung soll das Paketvermittlungsnetz bezüglich der Pakete über zwei Prioritätsklassen verfügen. Dies ist beispielsweise in DATAPAC [Sproule / Mellor (1981)] der Fall.

Die Prioritätsklasse 1 ist vorgesehen für die Netzsteuerung, Realzeitapplikationen und Dialogbetrieb, die eine kurze Netzdurchlaufzeit erfordern. Pakete des Stapelbetriebs können mit 2. Priorität abgewickelt werden, weil ihre Durchlaufzeiten keinen strengen Kriterien unterliegen. In diesem Zusammenhang sei hier bemerkt, daß die Zeitbegrenzungen für die Quittierungszeit (Time-Out) für beide Prioritäten verschieden gesetzt werden. Die Priorität eines jeden Paketes wird im Paketkopf mitgeführt, so daß die zugehörige Prioritätsklasse jederzeit identifiziert werden kann.

Zur Implementierung der FG-BG-Strategie soll jeder Netzknoten neben dem normalen Vermittlungspeicher (Foreground) über weitere Speichermöglichkeiten in einem Hintergrundspeicher (Background) verfügen können. Dieser Hintergrundspeicher ist bestimmt für die Auslagerung von Paketen der Prioritätsklasse 2, für die längere Netzdurchlaufzeiten in Kauf genommen werden können. Somit kann die Vermittlung dieser Pakete ohne gravierende Einbußen auf Perioden mit einer niedrigeren Verkehrsbelastung verlegt werden. Durch diese dynamische Speicherverwaltung wird der Vermittlungsspeicher im Bedarfsfalle für hochprioritäre Pakete freigegeben, indem die zurückgestellten Pakete der 2. Prioritätsklasse auf einen Hintergrundspeicher ausgelagert werden.

In einer Verkehrsumgebung mit hohem Anteil an Stapelverkehr soll von vornherein ein Teil des Vermittlungsspeichers für Pakete der 1. Priorität reserviert werden. Auf diese Weise wird verhindert, daß die Stapelpakete den gesamten Speicherbereich belegen können und dadurch insbesondere die lebenswichtigen Netzsteuerungspakete abgewiesen werden müssen (vgl. Abschnitt 5.4). Die Speicherbelegungseinschränkungen sind Software-Parameter und können adaptiv der zu erwartenden Verkehrsumgebung angepaßt werden.

7.2 Modellbeschreibung

Zur Untersuchung der Eigenschaften dieser Strategie wird das im Bild 7.1 dargestellte Verkehrsmodell betrachtet. Es besteht aus einem Warteschlangenmodell mit zwei nichtunterbrechenden Prioritätsklassen, die von einer einzigen Bedienungseinheit mit

unterschiedlichen Bedienungsraten abgefertigt werden. Die Anforderung der 1. Priorität werden im weiteren als 1-Anforderungen, die der 2. Priorität als 2-Anforderungen bezeichnet.

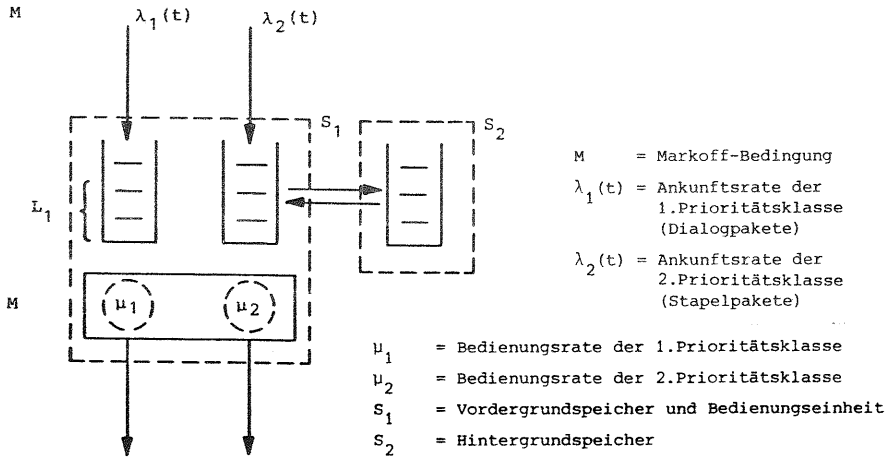


Bild 7.1: Verkehrsmodell: Foreground-Background Strategie.

Die für jede Klasse zugängliche Systemkapazität ist durch drei Parameter festgelegt:

- S_1 : **Kapazität im Vordergrundspeicher, der von beiden Prioritätsklassen in Anspruch genommen werden kann. Sie enthält auch die Anforderung in der Bedienungseinheit.**
- S_2 : **Kapazität im Hintergrundspeicher, der nur den 2-Anforderungen zugänglich ist.**
- L_1 : **Kapazität im Vordergrundspeicher, der für 1-Anforderungen reserviert ist.**

Das Verkehrsmodell stellt die gesamte Systemkapazität in einem Netzknoten dar, wobei die Pakete den Netzknoten mit einer prioritätsabhängigen Rate μ_1 bzw. μ_2 verlassen. Es wird angenommen, daß die interne Verarbeitungszeit für die Speicherverwaltung gegenüber den Paketübertragungszeiten vernachlässigt werden kann.

Die Ankunftsprozesse für die Anforderungen werden als Poisson-Prozesse modelliert. Dabei wird die zeitabhängige Ankunftsrate der 1-Anforderungen mit $\lambda_1(t)$, diejenige der 2-Anforderungen mit $\lambda_2(t)$ bezeichnet. Die zufälligen Bedienungszeiten seien negativ exponentiell verteilt mit dem Mittelwert $h_1 = 1/\mu_1$ bzw. $h_2 = 1/\mu_2$, abhängig von der Priorität.

Beschreibt man die Anzahl der 1-Anforderungen im System mit dem Zufallsvariable X_1 und entsprechend die Anzahl der 2-Anforderungen im System mit dem Zufallsvariable X_2, so werden in der FG-BG Überlastabwehrstrategie die Anforderungen nach folgenden Regeln vom System aufgenommen:

- 1-Anforderungen werden akzeptiert, wenn

$$X_1 + X_2 < S_1 + S_2 \quad \text{und} \quad \begin{cases} X_1 < S_1 - 1 & \text{, für eine 2-Anforderung in der Bedienungseinheit} \\ X_1 < S_1 & \text{, sonst.} \end{cases} \tag{7.1}$$

- 2-Anforderungen werden akzeptiert, wenn

$$X_1 + X_2 < S_1 + S_2 \quad \text{und} \quad X_2 < S_1 + S_2 - L_1 \quad . \tag{7.2}$$

7.3 Systemzustandsprozeß

7.3.1 Zustandsdiagramm

Grundlage für die Berechnung der stationären sowie transienten Zustandswahrscheinlichkeiten ist das Markoff-Zustandsdiagramm, dessen Struktur in Bild 7.2 dargestellt ist.

Jeder Zustand ist gekennzeichnet durch das Zahlentripel (i,j,k) mit

- i = Anzahl der 1-Anforderungen im System , $i = 0,\ldots,S_1$
- j = Anzahl der 2-Anforderungen im System , $j = 0,\ldots,S_1+S_2-L_1$
- k = Zustand der Bedienungseinheit
  - k = 0 : leeres System
  - k = 1 : Bedienung einer 1-Anforderung
  - k = 2 : Bedienung einer 2-Anforderung.

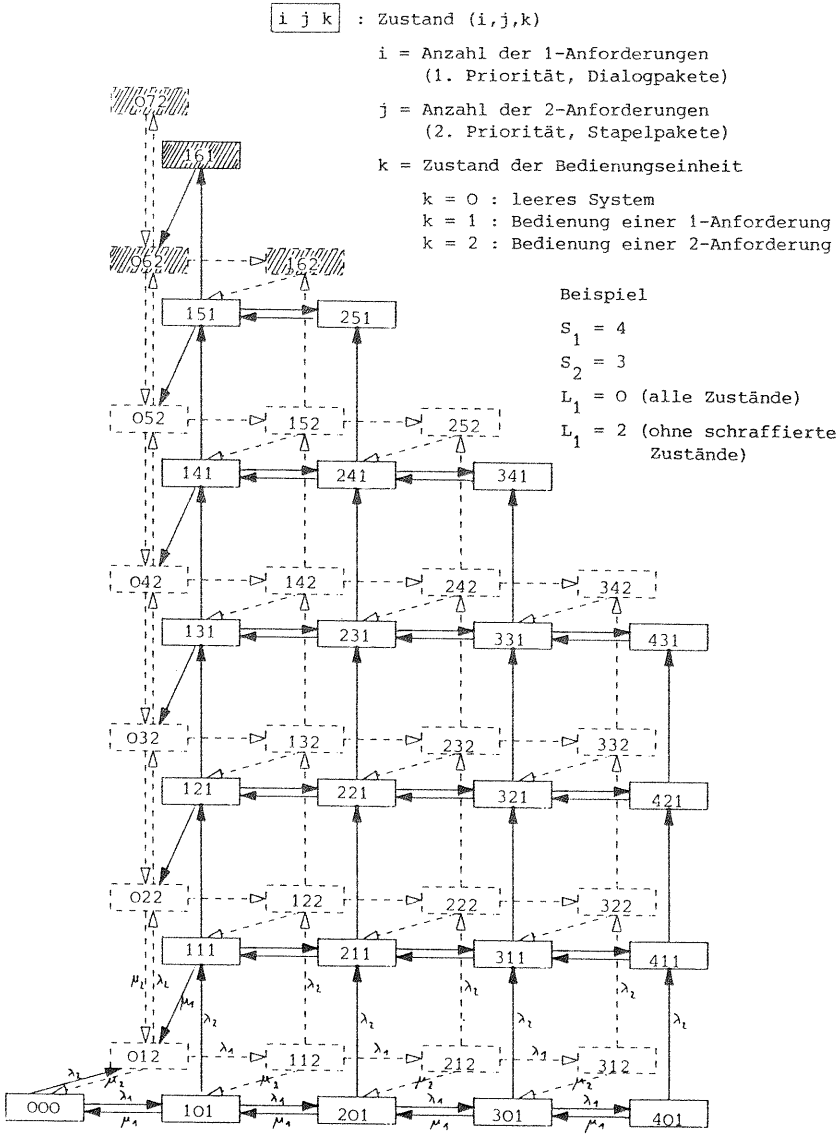


Bild 7.2: Zustandsdiagramm (Foreground-Background Strategie).

Der leere Systemzustand und alle Zustände mit einer 1-Anforderung in der Bedienungseinheit sind durchgezogen dargestellt. Alle Zustände mit Bedienung einer 2-Anforderung sind gestrichelt gezeichnet. Dies gilt auch für die zugehörigen Übergangspfeile. Da den 1-Anforderungen nur die Speicherkapazität S_1 (Vordergrundspeicher) zur Verfügung steht, wird das Zustandsdiagramm in der durchgezogenen Zustandsebene durch die Zustände $(S_1,j,1)$, $j = 0,\ldots S_2$, begrenzt. In der gestrichelten Zustandsebene bilden die Zustände $(S_1-1,j,2)$, $j = 1,\ldots,S_2+1$, die Grenze des Zustandsraumes, denn die bediente 2-Anforderung gehört ebenfalls zur Kapazität S_1 des Vordergrundspeichers. Sind i Anforderungen der 1. Priorität im System vorhandenen, so bleiben noch höchstens $S_1 + S_2 - \max(i,L_1)$ Speicherplätze für Anforderungen der 2. Priorität übrig. Dabei ist $\max(i,L_1)$ das Maximum von i und L_1. Im Zustandsdiagramm wird die teilweise gemeinsame Speichernutzung als diagonale Abgrenzung, die Speicherplatzreservierung für 1-Anforderung als schraffierter Bereich von unzugänglichen Zuständen berücksichtigt. Entsprechend dem Betriebsmodus für nichtunterbrechende Prioritäten findet beim Bedienungsende einer 2-Anforderung ein Klassenwechsel für die nächste Bedienung statt, solange 1-Anforderungen vorhanden sind (Wechsel von der gestrichelten Zustandsebene zur durchgezogenen Zustandsebene).

Warten keine 1-Anforderungenauf Bedienung, so wird die nächste 2-Anforderung bedient (linke Zustandskolonne in der gestrichelten Zustandsebene bzw. Zustände $(0,j,2)$, $j = 1,\ldots,S_1+S_2-L_1$) oder das System geht in den leeren Zustand über. Ein Wechsel von der durchgezogenen zur gestrichelten Zustandsebene erfolgt nur, wenn alle 1-Anforderungen bedient sind und noch 2-Anforderungen auf Bedienung warten (linke durchgezogene Zustandskolonne bzw. Zustände $(1,j,1)$, $j = 1,\ldots,S_1+S_2-L_1$).

7.3.2 Rekursive Berechnung der stationären Zustandswahrscheinlichkeiten

Zur Bestimmung der stationären Zustandswahrscheinlichkeiten kann die iterative Methode herangezogen werden. Aufgrund der Struktur dieses Zustandsdiagramms lassen sich die Zustandswahrscheinlichkeiten jedoch auch rekursiv bestimmen.

Im folgenden werden die einzelnen Lösungsschritte eines rekursiven Lösungsverfahrens abgeleitet und zum Schluß in Form von einem Algorithmus zusammengefaßt. Unter Berücksichtigung modellspezifischer Abänderungen hat dieser Algorithmus eine allgemeine Gültigkeit für einstufige Markoff-Warteschlangensysteme mit zwei nichtunterbrechenden Prioritäten und mit verschiedenen Einschränkungen bezüglich der Systemkapazität. Insbesondere gilt er auch für die Prioritätsmodelle in den Abschnitten 5.4 und 6. Im vorliegenden Modell wird nur der Fall ohne Reservierung für 1-Anforderungen ($L_1 = 0$) betrachtet.

Im weiteren werden sämtliche Zustände mit einer 1-Anforderung in der Bedienungseinheit als Priorität-1-Zustände bezeichnet. Alle Priorität-1-Zustände, die im Zustandsdiagramm zur gleichen Zeile gehören, werden Priorität-1-Zeile genannt. Analog gilt diese Bezeichnungsweise auch für die 2. Prioritätsklasse.

Zuerst wird die Notation $p_i^{j,k}$ für die Wahrscheinlichkeit, daß der Zustand (i,j,k) auftritt, eingeführt. Mit dieser Schreibweise läßt sich aus Bild 7.2 das nachfolgende System von rekursiven Gleichungen für die Zustandswahrscheinlichkeiten aufstellen. Die Zustandswahrscheinlichkeiten werden zeilenweise berechnet. Dazu wird jeweils der Zustand am rechten Rand des Zustandsdiagramms (maximaler Wert für Index i) als Startpunkt für die Rekursion ausgewählt.

Durch suksessive Anwendung von statistischen Gleichgewichtsbetrachtungen, beginnend mit dem äußerst rechts angeordneten Zustand, erhält man für die unterste Priorität-1-Zeile:

$$p_{S_1-1}^{0,1} = \frac{\mu_1+\lambda_2}{\lambda_1} \cdot p_{S_1}^{0,1}$$

$$p_i^{0,1} = \frac{\mu_1}{\lambda_1} \cdot p_{i+1}^{0,1} + \frac{\lambda_2}{\lambda_1} \cdot \sum_{n=S_1}^{i+1} p_n^{0,1} - \frac{\mu_2}{\lambda_1} \cdot \sum_{n=S_1-1}^{i+1} p_n^{1,2} \quad , \; i = S_1-2,\ldots,0 \qquad (7.3)$$

wobei die Identität $p_0^{0,0} \equiv p_0^{0,1}$ gilt.

Entsprechend gilt für die unterste Priorität-2-Zeile:

$$p_i^{1,2} = \frac{\mu_2+\lambda_2}{\lambda_1} \cdot \sum_{n=S_1-1}^{i+1} p_n^{1,2} \quad , \; i = S_1-2,\ldots,0 \; . \qquad (7.4)$$

Unter Einbeziehung des maximalen Wertes für Index i:

$$m = \begin{cases} S_1 & \text{für} \quad j < S_2 \\ S_1 + S_2 - j & \text{für} \quad j \geq S_2 \end{cases} \quad , \; j = 1,\ldots,S_1+S_2 \qquad (7.5)$$

läßt sich für die weiteren Priorität-1-Zeilen schreiben:

$$p_{m-1}^{j,1} = \frac{\mu_1+\lambda_2}{\lambda_1} \cdot p_m^{j,1} - \frac{\lambda_2}{\lambda_1} \cdot p_m^{j-1,1}$$

$$p_i^{j,1} = \frac{\mu_1}{\lambda_1} \cdot p_{i+1}^{j,1} + \frac{\lambda_2}{\lambda_1} \cdot \sum_{n=m}^{i+1} p_n^{j,1} - \frac{\lambda_2}{\lambda_1} \cdot \sum_{n=m}^{i+1} p_n^{j-1,1} - \frac{\mu_2}{\lambda_1} \cdot \sum_{n=m-1}^{i+1} p_n^{j+1,2} \qquad (7.6)$$

$$, \; i = m-2,\ldots,1 \quad .$$

In ähnlicher Weise bekommt man für die weiteren Priorität-2-Zeilen:

$$p_i^{j+1,2} = \frac{\mu_2+\lambda_2}{\lambda_1} \cdot \sum_{n=m-1}^{i+1} p_n^{j+1,2} - \frac{\lambda_2}{\lambda_1} \cdot \sum_{n=m-1}^{i+1} p_n^{j,2} \quad , \; i = m-2,\ldots,0 \quad . \qquad (7.7)$$

Schließlich gilt für die letzten zwei Zustände:

$$p_1^{S-1,1} = \frac{\lambda_2}{\mu_1} \cdot p_1^{S-2,1}$$

$$p_0^{S,2} = \frac{\lambda_2}{\mu_2} \cdot p_0^{S-1,2} \qquad , \; S = S_1 + S_2 \qquad (7.8)$$

Zusammenfassend gilt also, daß aus den als bekannt vorausgesetzten Zustandswahrscheinlichkeiten am rechten Rand des Zustandsdiagramms die restlichen Zustandswahrscheinlichkeiten rekursiv berechnet werden können. Dies gilt sowohl für die Priorität-1-Zustände, als auch für die Priorität-2-Zustände.

Wie bereits im Abschnitt 4.4.2 erwähnt wurde, werden im Gegensatz zum Originalverfahren nach [Herzog/Woo/Chandy (1975)] die notwendigen a-priori-Wahrscheinlichkeiten nicht erst am Ende, sondern bereits im Laufe des rekursiven Berechnungsvorganges eliminiert. Dies führt zu einer besseren Speicherplatz- und Rechenzeiteffizienz, so daß auch umfangreichere Zustandsräume kostengünstig berechnet werden können.

Im nachfolgenden Algorithmus ist die Anzahl der erforderlichen Rekursionsstartpunkte abhängig von dem momentan durchgeführten Berechnungsschritt. Dabei sind höchstens drei Rekursionsstartpunkte gleichzeitig aktiv, so daß in diesem Fall die entsprechende Linearkombination für die Zustandswahrscheinlichkeiten lautet:

$$P_i^{j,k} = A_i^{j,k} \cdot P_X^1 + B_i^{j,k} \cdot P_X^2 + C_i^{j,k} \cdot P_X^3 \quad . \tag{7.9}$$

Das Einführen und Eliminieren von a-priori-Wahrscheinlichkeiten im Laufe des Algorithmus erfolgt auf folgende Weise (Bild 7.2):

1) Als Basis für die gesamte Rekursion wird der Anfangspunkt $P_{S_1-1}^{1,2} = P_X^1$ genommen. Diese a-priori-Wahrscheinlichkeit wird erst am Ende durch Normierung bestimmt.
   Alle Zustandswahrscheinlichkeiten, die also nur noch von diesem Normierungsfaktor P_X^1 abhängen, werden durch eine obere Schlange charakterisiert: $\breve{P}_i^{j,k}$.

2) Die Berechnung der untersten Priorität-2-Zeile kann direkt mit P_X^1 durchgeführt werden.

3) Die Berechnung der untersten Priorität-1-Zeile erfordert die Einführung eines zusätzlichen Rekursionsstartpunktes $P_{S_1}^{0,1} = P_X^2$. Er kann jedoch, bevor zum nächsten Zeilenpaar (Priorität 1 und Priorität 2) übergegangen wird, eliminiert werden. Dies geschieht durch zwei unabhängige Berechnungen für die Zustandswahrscheinlichkeit $P_0^{1,2}$ und anschließendes Gleichsetzen:

   - die Berechnung nach Gl.(7.4) liefert mit Hilfe einer statistischen Gleichgewichtsbetrachtung für die unterste Priorität-2-Zeile, wie dies in Bild 7.3a dargestellt ist, den linken Teil der Gl.(7.10),

   - die Berechnung nach Gl.(7.3) liefert mit Hilfe einer entsprechenden Betrachtung für die unterste Priorität-1-Zeile, wie dies in Bild 7.3b dargestellt ist, den rechten Teil der Gl.(7.10),

   - das Gleichsetzen der beiden algebraischen Ausdrücke für $P_0^{1,2}$ ergibt eine lineare Gleichung zur Bestimmung von P_X^2 als Funktion von P_X^1

$$\breve{P}_0^{1,2} \cdot P_X^1 = A_0^{1,2} \cdot P_X^1 + B_0^{1,2} \cdot P_X^2 \quad . \tag{7.10}$$

   Nach der Elimination dieses Rekursionsstartpunktes P_X^2 lassen sich sämtliche, bis jetzt betrachteten Zustandswahrscheinlichkeiten als Funktion von P_X^1 ausdrücken.

4) Die Berechnung jedes nachfolgenden Zeilenpaares erfordert jeweils zwei zusätzliche Rekursionsstartpunkte. Wird die j-te Zeile betrachtet und hat der Index i den maximalen Wert m, so gilt:

$$P_{m-1}^{j+1,2} = P_X^2 \quad \text{und} \quad P_m^{j,1} = P_X^3 \quad . \tag{7.11}$$

   Dabei wird für eine Priorität-2-Zeile nur P_X^2 benötigt. Die pro Zeilenpaar eingeführten, zusätzlichen a-priori-Zustandswahrscheinlichkeiten werden jeweils mit Hilfe von zwei Relationen eliminiert. Zur Erläuterung sind diese zwei Beziehungen für j = 2 in den Bildern 7.3c bzw. 7.3d dargestellt.

   - Für die erste Beziehung (Bild 7.3c) wird der in den vorangegangenen Schritten bereits berechnete Teil des Zustandsraumes zusammengefaßt, so daß bei Anwendung der statistischen Gleichgewichtsbedingung alle internen Übergänge wegfallen.

$$\lambda_2 \cdot \left[\sum_{n=m-1}^{0} \breve{P}_n^{j,2} + \sum_{n=m}^{1} \breve{P}_n^{j-1,1} \right] = \underbrace{\mu_1 \cdot P_1^{j,1} + \mu_2 \cdot P_0^{j+1,2}}_{\text{momentan betrachtetes Zeilenpaar}} \tag{7.12}$$

   - Für die zweite Beziehung (Bild 7.3d) werden sämtliche, bis jetzt berechneten Priorität-1-Zustände zusammengefaßt, die

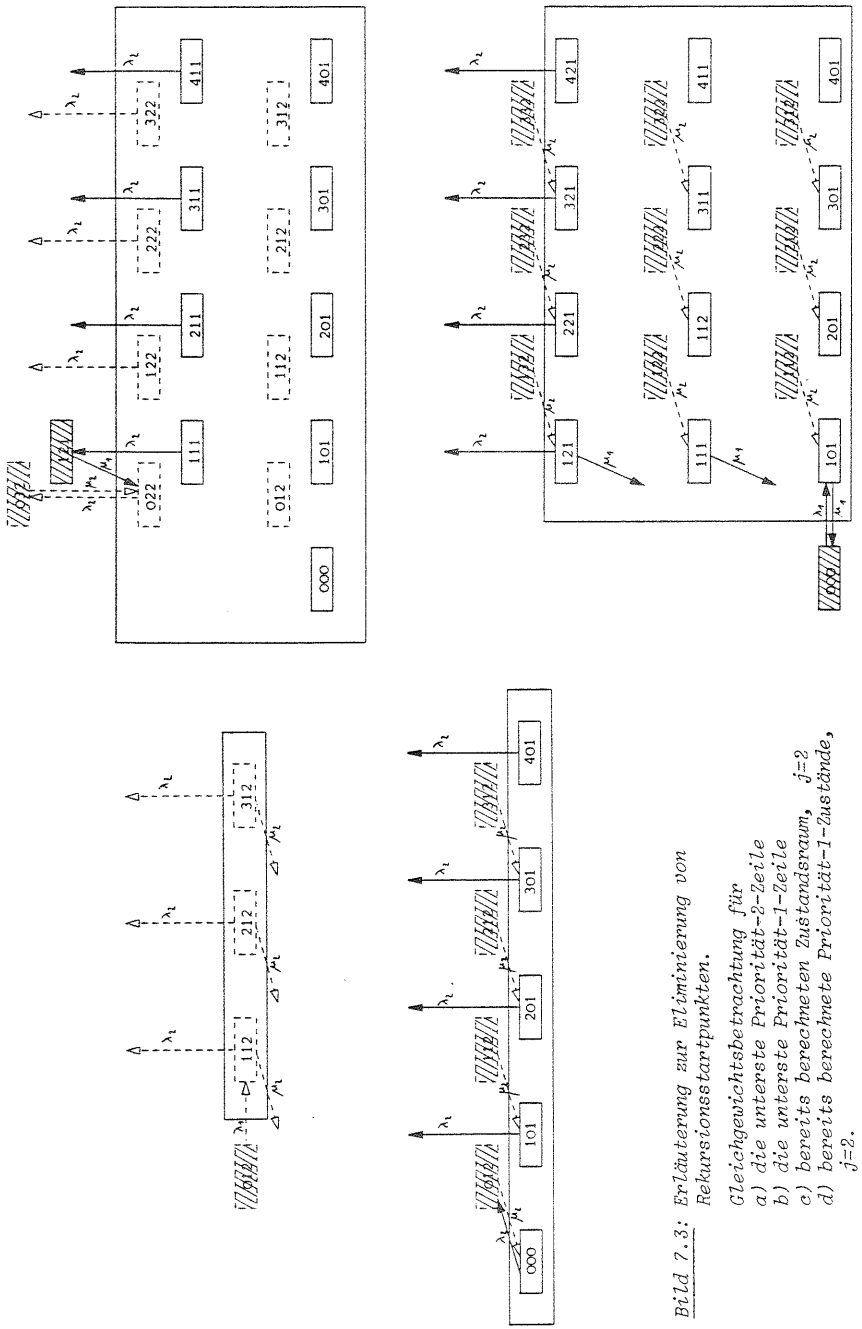


Bild 7.3: Erläuterung zur Eliminierung von Rekursionsstartpunkten.
Gleichgewichtsbetrachtung für
a) die unterste Priorität-2-Zeile
b) die unterste Priorität-1-Zeile
c) bereits berechneten Zustandsraum, j=2
d) bereits berechnete Priorität-1-Zustände, j=2.

mit den schraffiert dargestellten Priorität-2-Zuständen und dem Zustand (O,O,O) in statistischem Gleichgewicht sein müssen. Die Übergänge zwischen den Priorität-1-Zuständen fallen dabei weg. Somit gilt:

$$\underbrace{\lambda_2 \cdot \sum_{n=m}^{1} P_n^{j,1} + \mu_1 \cdot P_1^{j,1}}_{\text{momentan betrachtetes Zeilenpaar}} + \mu_1 \cdot \sum_{k=O}^{j-1} \tilde{P}_1^{k,1} = \underbrace{\mu_2 \cdot \sum_{n=m-1}^{1} P_n^{j+1,2}}_{\text{momentan betrachtetes Zeilenpaar}} + \mu_2 \cdot \sum_{k=1}^{j} \sum_{n=m-1}^{1} \tilde{P}_n^{k,2} + \lambda_1 \cdot \hat{P}_O^{O,O} \qquad (7.13)$$

- Für die Durchführung der Elimination wird das algebraische Ergebnis von Gl.(7.12), das vom Rekursionsstartvektor $(P_X^1, P_X^2, P_X^3) = (V,O,O)$ abhängt, als α bezeichnet. Zum Vektor (O,V,O) gehört β und γ erhält man mit Hilfe des Vektors (O,O,V). Die entsprechende Bezeichnung $\tilde{\alpha}$, $\tilde{\beta}$ und $\tilde{\gamma}$ gelten für die Gl.(7.13). Damit erhält man zwei unabhängige lineare Gleichungen aus denen P_X^2 sowie P_X^3 als Funktion von P_X^1 ermittelt werden können:

$$\alpha \cdot P_X^1 + \beta \cdot P_X^2 + \gamma \cdot P_X^3 = O$$

$$\tilde{\alpha} \cdot P_X^1 + \tilde{\beta} \cdot P_X^2 + \tilde{\gamma} \cdot P_X^3 = O \qquad (7.14)$$

Nach diesem Eliminationsschritt können auch die Zustandswahrscheinlichkeiten des momentan betrachteten Zeilenpaares als Funktion von P_X^1 ausgedrückt werden.

Die Zusammenfassung der einzelnen Lösungsschritte führt zu folgendem Algorithmus:

1) Initialisierung des ersten Startpunktes $\tilde{P}_{S_1-1}^{1,2} = P_X^1 = V$ und die direkte Berechnung der Zustandswahrscheinlichkeiten $\tilde{P}_i^{1,2}$ der untersten Priorität-2-Zeile (j = 1) nach Gl.(7.4), $i = S_1-2,\ldots,O$.

2) Einführung eines zweiten Startpunktes $P_{S_1}^{O,1} = P_X^2$ und nachfolgende Berechnung der Koeffizienten $A_i^{O,1}$ bzw. $B_i^{O,1}$ der untersten Priorität-1-Zeile (j = O) nach Gl.(7.3), wobei die bereits berechneten Zustandswahrscheinlichkeiten $\tilde{P}_i^{1,2}$ direkt berücksichtigt werden, $i = S_1-1,\ldots,O$.

3) Elimination von P_X^2 nach Gl.(7.10) und Bestimmung der Zustandswahrscheinlichkeiten $\tilde{P}_i^{O,1}$ nach Gl.(7.9), $i = S_1-1,\ldots,O$.

4) Bestimmung des maximalen Wertes m für Index i nach Gl.(7.5) für das nächste Zeilenpaar: Priorität 1 (Zeile j) bzw. Priorität 2 (Zeile j+1).

5) Einführung eines neuen zweiten Startpunktes $P_{m-1}^{j+1,2} = P_X^2$ und Berechnung der Koeffizienten $A_i^{j+1,2}$ bzw. $B_i^{j+1,2}$ der Priorität-2-Zeile j+1 nach Gl.(7,7), $i = m-2,\ldots,O$.

6) Einführung eines (neuen) dritten Startpunktes $P_m^{j,1} = P_X^3$ und Berechnung der Koeffizienten $A_i^{j,1}$, $B_i^{j,1}$ bzw. $C_i^{j,1}$ der Priorität-1-Zeile j nach Gl.(7.6), $i = m-1,\ldots,1$.

7) Elimination der eingeführten Startpunkte P_X^2 und P_X^3 nach Gl.(7.14) und Bestimmung der Zustandswahrscheinlichkeiten $\tilde{P}_i^{j,1}$ $(i = m-1,\ldots,1)$ und $\tilde{P}_i^{j+1,2}$ $(i = m-2,\ldots,O)$ nach Gl.(7.9).

8) Inkrementierung von j und falls $j < S_1+S_2-1$ Wiederholung ab Schritt 4.

9) Berechnung der beiden letzten Zustandswahrscheinlichkeiten nach Gl.(7.8).

10) Bestimmung von P_X^1 durch Normierung der Zustandsverteilung und nachfolgende Berechnung der normierten Zustandswahrscheinlichkeiten.

Ein sehr ähnlicher Algorithmus läßt sich auch mit Hilfe von statistischen Gleichgewichtsbedingungen bezüglich einzelner Zustände formulieren. Dadurch ändern sich lediglich die Gleichungen, das Ablaufschema selbst bleibt unverändert. Zur Eliminierung des zweiten Rekursionsstartpunktes P_X^2 für das unterste Zeilenpaar wird dann die Gleichgewichtsgleichung für die Zustandswahrscheinlichkeit $P_O^{O,O}$ benötigt. Die Eliminierung der beiden zusätzlichen Rekursionsstartpunkte für die weiteren Zeilenpaare erfolgt mit Hilfe der entsprechenden Gleichungen für $P_O^{j,2}$ und $P_1^{j,1}$, mit $j = 1,\ldots,S_1+S_2-1$.

7.3.3 Berechnung der transienten Zustandswahrscheinlichkeiten

Die transienten Zustandswahrscheinlichkeiten $P_i^{j,k}(t)$ werden berechnet durch die numerische Lösung der gekoppelten Differentialgleichungen für die Zustandswahrscheinlichkeiten (vgl. Abschnitt 4.4.3). Dieses System von Differentialgleichungen läßt sich mit Hilfe des Zustandsdiagramms (Bild 7.2) aufstellen. Die Regeln dazu sind im Abschnitt 4.2.3 angegeben.

Als Beispiel sind die Differentialgleichungen für die Zustände der untersten Zeile nachstehend aufgeführt:

$$\frac{d}{dt} P_O^{O,O}(t) = -[\lambda_1(t)+\lambda_2(t)] \cdot P_O^{O,O}(t) + \mu_1 \cdot P_1^{O,1}(t) + \mu_2 \cdot P_O^{1,2}(t)$$

$$\frac{d}{dt} P_i^{O,1}(t) = -[\lambda_1(t)+\lambda_2(t)+\mu_1] \cdot P_i^{O,1}(t) + \lambda_1(t) \cdot P_{i-1}^{O,1}(t) + \mu_1 \cdot P_{i-1}^{O,1}(t)$$
$$+ \mu_2 \cdot P_i^{1,2}(t) \qquad , i = 1,\ldots,S_1-1$$

$$\frac{d}{dt} P_{S_1}^{O,1}(t) = -[\lambda_2(t)+\mu_1] \cdot P_{S_1}^{O,1}(t) + \lambda_1(t) \cdot P_{S_1-1}^{O,1}(t) \qquad (7.15)$$

7.4 Durchlaufprozeß

Die mittlere Durchlaufzeit einer zum Zeitpunkt s eintreffenden 1-Anforderung $t_{F1}(s)$ läßt sich bei einer FIFO-Abfertigungsstrategie direkt aus dem angetroffenen Systemzustand ermitteln, denn in diesem Fall wird die betrachtete 1-Anforderung nicht mehr von später eintreffenden Anforderungen beeinflußt. Werden jedoch andere Abfertigungsstrategien angewendet oder wird die entsprechende mittlere Durchlaufzeit einer 2-Anforderung $t_{F2}(s)$ gesucht, so muß der Durchlaufprozeß betrachtet werden.

Zur Berechnung von $t_{F2}(s)$ wird ein Zustand (i,j,k) des Durchlaufprozesses bezüglich einer 2-Anforderung wie folgt definiert:

i = Anzahl der 1-Anforderungen, die vor der Testanforderung der 2. Prioritätsklasse bedient werden. Sie sind entweder beim Eintreffen der betrachteten 2-Anforderung bereits anwesend oder treffen während ihrer Wartezeit später ein.

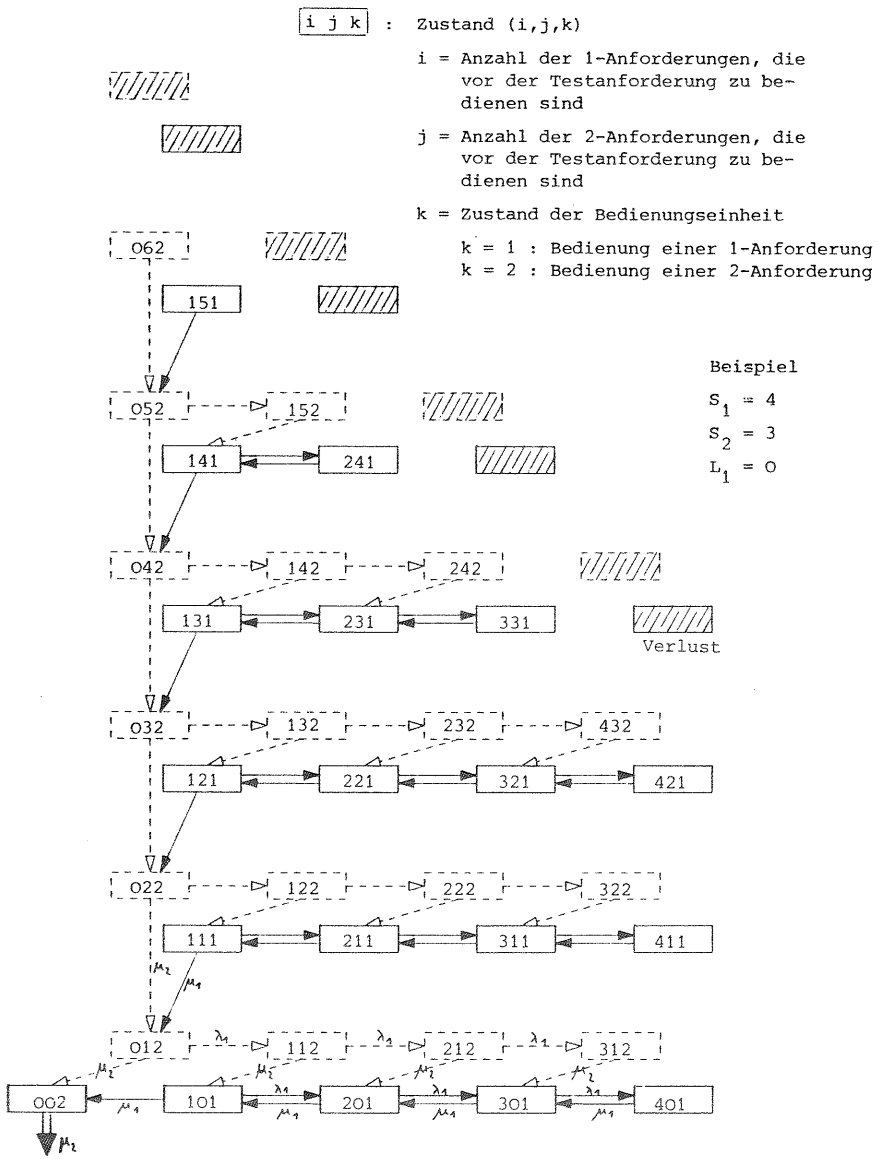


Bild 7.4: Durchlaufzustandsdiagramm für eine Testanforderung der 2. Priorität (Foreground-Background Strategie).

j = Anzahl der 2-Anforderungen, die vor der Testanforderung der 2. Prioritätsklasse bedient werden. Sie werden beim Eintreffen der betrachteten 2-Anforderung im System vorgefunden.

k = Zustand der Bedienungseinheit
k = 1 : Bedienung einer 1-Anforderung
k = 2 : Bedienung einer 2-Anforderung bzw. der Testanforderung.

Bild 7.4 zeigt das Durchlaufzustandsdiagramm für 2-Anforderungen mit FIFO-Abfertigungsstrategie. Für den Fall, daß die Testanforderung der 2. Prioritätsklasse die schraffierten Zustände antrifft (volles System), wird sie abgewiesen. In allen anderen Zuständen wird der Durchlaufprozeß begonnen und endet nach der Bedienung der Testanforderung. Ihre Bedienung wird symbolisiert durch den Zustand (0,0,2). Dieser "leere" Systemzustand bedeutet, daß die Testanforderung nicht mehr von anderen Anforderungen beeinflußt werden kann. Wie aus dem Zustandsdiagramm hervorgeht, kann die Durchlaufzeit einer Testanforderung durch später eintreffende 1-Anforderungen, die bevorzugt bedient werden, verzögert werden. Neue 1-Anforderungen können jedoch nicht mehr alle Zustände erreichen, weil ein Speicherplatz bereits von der Testanforderung belegt worden ist. Die nicht erreichbaren Systemzustände sind genau die schraffierten Zustände. Sie sind deshalb vom Durchlaufzustandsdiagramm ausgeschlossen.

Als Beispiel sei eine Testanforderung der 2. Prioritätsklasse betrachtet, die den Zustand (1,2,2) antrifft. Durch neue 1-Anforderungen werden die rechts liegenden Zustände (i,2,2), $i = 2,\ldots,S_1-1$ nacheinander durchlaufen. Treffen weitere 1-Anforderungen ein, so müssen sie abgewiesen werden. Bei Bedienungsende der 2-Anforderung fängt ein Bedienungszyklus von 1-Anforderungen an, wobei alle 1-Anforderungen, die während dieser Zeit eintreffen, ebenfalls abgefertigt werden: Zustände (i,1,1), $i = 1,\ldots,S_1$. Danach wird eine weitere 2-Anforderung bedient, so daß der Durchlaufprozeß sich jetzt im Zustand (0,1,2) befindet. Falls während dieser Bedienungszeit 1-Anforderungen eintreffen, überholen auch sie die Testanforderung. Zuletzt erfolgt dann die Bedienung der betrachteten 2-Anforderungen: Zustand (0,0,2).

Mathematisch wird der Durchlaufprozeß durch die Kolmogoroff-Rückwärts-Differentialgleichungen für die bedingten komplementären Durchlaufzeitfunktionen $f_i^{j,k}(s,t)$ beschrieben. Diese Rückwärts-Differentialgleichungen lassen sich nach den im Abschnitt 4.2.4 zusammengestellten Regeln aus dem Durchlaufzustandsdiagramm ableiten. Dies wird hier für die unterste Zustandszeile gezeigt:

$$\frac{d}{ds} f_O^{O,2}(s,t) = -\mu_2 \cdot f_O^{O,2}(s,t)$$

$$\frac{d}{ds} f_1^{O,1}(s,t) = -[\mu_1+\lambda_1(s)] \cdot f_1^{O,1}(s,t) + \lambda_1(s) \cdot f_2^{O,1}(s,t) + \mu_1 \cdot f_O^{O,2}(s,t)$$

$$\frac{d}{ds} f_i^{O,1}(s,t) = -[\mu_1+\lambda_1(s)] \cdot f_i^{O,1}(s,t) + \lambda_1(s) \cdot f_{i+1}^{O,1}(s,t) + \mu_1 \cdot f_{i-1}^{O,1}(s,t)$$

$$i = 2,\ldots,S_1-1$$

$$\frac{d}{ds} f_{S_1}^{O,1}(s,t) = -\mu_1 \cdot f_{S_1}^{O,1}(s,t) + \mu_1 \cdot f_{S_1-1}^{O,1}(s,t) \quad . \tag{7.16}$$

Die numerische Lösung dieses Systems von linearen Differentialgleichungen liefert für jeden Zustand (i,j,k) die entsprechende bedingte komplementäre Durchlaufzeitverteilungsfunktion $f_i^{j,k}(s,t)$, so daß durch Gewichtung mit den Antreffwahrscheinlichkeiten gemäß Gl.(4.26) die Durchlaufzeitverteilungsfunktion bezüglich einer zum Zeitpunkt s eintreffenden 2-Anforderung erhalten wird:

$$F_2^c(s,t) = \sum_{\notin H} p_i^{j,k}(s) \cdot f_i^{j,k}(s,t) \quad , \quad \text{H: Zustände, die zu Verlust führen.} \tag{7.17}$$

Die mittlere Durchlaufzeit $t_{F2}(s)$ erhält man durch numerische Integration der Durchlaufzeitverteilungsfunktion.

7.5 Charakteristische Verkehrsgrößen

Zur Beurteilung der FG-BG Strategie werden die folgenden charakteristischen Verkehrsgrößen definiert:

- Verlustwahrscheinlichkeit für 1-Anforderungen:

$$B_1(t) = \underbrace{\sum_{\text{Bedingung 1}} [p_i^{j,1}(t)+p_i^{j,2}(t)]}_{\text{gemeinsamer Speicher}} + \underbrace{\sum_{j=1}^{S_2} [p_{S_1}^{j-1,1}(t)+p_{S_1-1}^{j,2}(t)]}_{\text{Vordergrund Speicher}} \tag{7.18}$$

- Verlustwahrscheinlichkeit für 2-Anforderungen:

$$B_2(t) = \underbrace{\sum_{\text{Bedingung 1}} [p_i^{j,1}(t)+p_i^{j,2}(t)]}_{\text{gemeinsamer Speicher}} + \underbrace{\sum_{\text{Bedingung 2}} p_i^{j,1}(t) + \sum_{\text{Bedingung 3}} p_i^{j,2}(t)}_{\text{Speicherreservierung}} \tag{7.19}$$

wobei mit $S = S_1+S_2$ gilt

Bedingung 1 : $i+j=S$ und $j \le S-L_1$ (7.20)

Bedingung 2 : $L_1 > 0$ und $j = S-L_1$ und $1 \le i < S-j$

Bedingung 3 : $L_1 > 0$ und $j = S-L_1$ und $0 \le i < S-j$

- Mittlere Systembelastung durch 1-Anforderungen:

$$E[X_1(t)] = \sum_i \sum_j i \cdot [p_i^{j,1}(t)+p_i^{j,2}(t)] \tag{7.21}$$

- Mittlere Systembelastung durch 2-Anforderungen:

$$E[X_2(t)] = \sum_i \sum_j j \cdot [p_i^{j,1}(t)+p_i^{j,2}(t)] \tag{7.22}$$

- Durchsatz von 1-Anforderungen:

$$D_1(t) = \mu_1 \cdot \sum_i \sum_j p_i^{j,1}(t) \tag{7.23}$$

- Durchsatz von 2-Anforderungen:

$$D_2(t) = \mu_2 \cdot \sum_i \sum_j p_i^{j,2}(t) \tag{7.24}$$

- Mittlere Durchlaufzeit für eine zum Zeitpunkt s eintreffende, akzeptierte 1-Anforderung:

  sie besteht aus drei Komponenten:
  - die Bedienung der Testanforderung selbst,
  - die Restbedienungszeit einer 2-Anforderung, falls beim Eintreffen der Testanforderung sich eine 2-Anforderung in der Bedienungseinheit befindet; die Wahrscheinlichkeit hierfür wird ermittelt durch Summation der Zustandswahrscheinlichkeiten aller Zustände (i,j,2), die nicht zum Verlust der Testanforderung führen (i,j$\notin$H). Bedingt durch die Markoff-Annahme ist die Restbedienungszeit ebenfalls negativ exponentiell verteilt und somit muß auch für die Restbedienung die mittlere Bedienungszeit h_2 angesetzt werden; da akzeptierte 1-Anforderungen betrachtet werden, wird diese Komponente noch mit der Verlustwahrscheinlichkeit $B_1(s)$ korrigiert,
  - die Gesamtbedienungszeit derjenigen 1-Anforderungen, die beim Eintreffen der Testanforderung bereits vorhanden sind. Die angetroffene mittlere Systembelastung von 1-Anforderungen, die vorher zu bedienen ist, bekommt man durch Bildung des Erwartungswertes, wobei nur Zustände, die nicht zum Verlust führen, berücksichtigt werden (i,j$\notin$H); weil eine bedingte mittlere Durchlaufzeit betrachtet wird, muß auch diese Komponente mit der Verlustwahrscheinlichkeit $B_1(s)$ korrigiert werden,

$$t_{F1}(s|\text{akzeptiert}) = \underbrace{h_1}_{\substack{\text{Bedienung}\\ \text{der}\\ \text{Testanfor-}\\ \text{derung}}} + \frac{1}{1-B_1(s)} \cdot \Bigg[\underbrace{h_2 \cdot \sum_i \sum_j P_i^{j,2}(s)}_{\substack{\text{Restbedienung}\\ \text{einer 2-Anfor-}\\ \text{derung}}} + \underbrace{h_1 \cdot \sum_i \sum_j i \cdot [P_i^{j,1}(s) + P_i^{j,2}(s)]}_{\substack{\text{Bedienung der angetroffenen}\\ \text{1-Anforderungen}}} \Bigg]$$

$$, \; i,j \notin H \qquad (7.25)$$

H: alle Zustände, die zum Verlust führen

- Mittlere Durchlaufzeit für eine zum Zeitpunkt s eintreffende, akzeptierte 2-Anforderung:

  sie läßt sich durch numerische Integration aus der entsprechenden komplementären Durchlaufzeitverteilungsfunktion Gl. (7.17) ermitteln; da eine bedingte mittlere Durchlaufzeit betrachtet wird, muß das Integrationsergebnis noch mit der Verlustwahrscheinlichkeit $B_2(s)$ korrigiert werden,

$$t_{F2}(s|\text{akzeptiert}) = \frac{1}{1-B_2(s)} \cdot \int_{\tau=s}^{\infty} F_2^C(s,\tau) \cdot d\tau \qquad (7.26)$$

7.6 Numerische Ergebnisse

Numerische Beispiele sollen nun die Verkehrseigenschaften der FG-BG-Strategie demonstrieren.

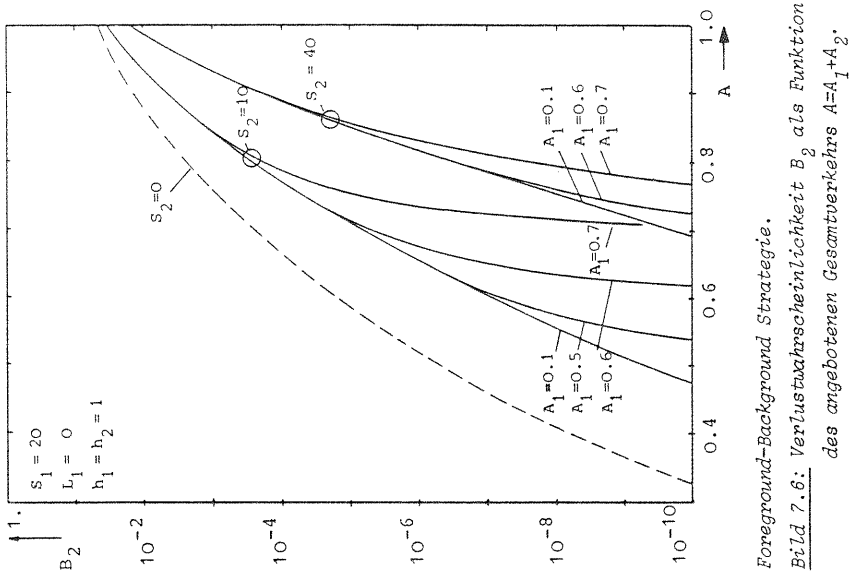
7.6.1 Stationäre Ergebnisse

Obwohl die FG-BG-Strategie dazu dient, kurze Überlastsituationen zu überbrücken und deshalb auch dynamisch zu untersuchen ist, werden zur Orientierung und zur Festlegung eines geeigneten Betriebspunktes zuerst stationäre Ergebnisse diskutiert.

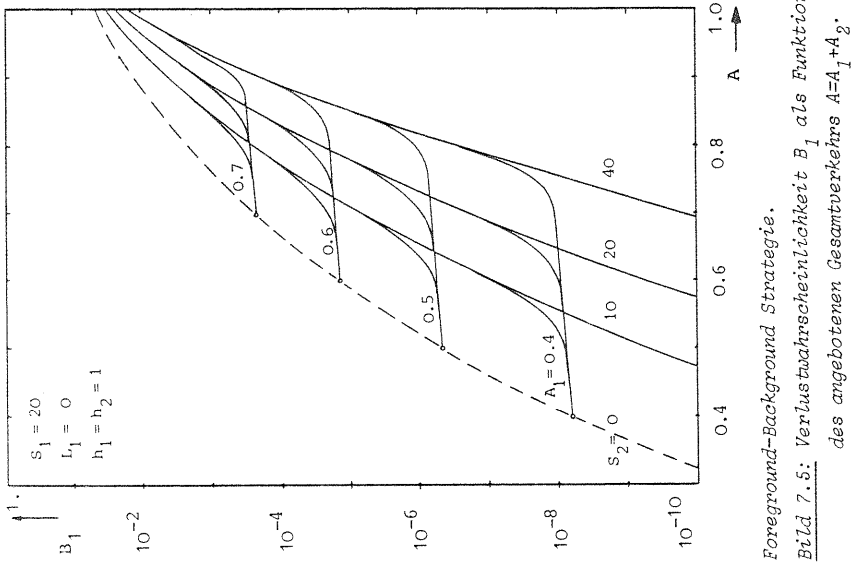
Betrachtet wird ein Warteschlangensystem mit einem Vordergrundspeicher $S_1 = 20$. Es werden keine Speicherplätze für 1-Anforderungen reserviert d.h. $L_1 = 0$. Ferner gilt für die mittlere Bedienungszeit der beiden Prioritätsklassen $h_1 = h_2 = 1$.

a) Verlustwahrscheinlichkeit für 1-Anforderungen

Bild 7.5 zeigt die Verlustwahrscheinlichkeit B_1 in Abhängigkeit des angebotenen Gesamtverkehrs $A = \lambda_1 \cdot h_1 + \lambda_2 \cdot h_2$. Die gestrichelte Kurve gilt für ein System mit einem gemeinsamen Speicher S_1 ohne die Möglichkeit zur Auslagerung von zurückgestellten 2-Anforderungen. Wird in diesem Falle (zum Beispiel bei einem festen Verkehrswert $A_1 = 0.4$) das Angebot der 2. Prioritätsklasse erhöht, so nimmt auch die Verlustwahrscheinlichkeit B_1 gemäß dieser gestrichelten Kurve zu. Können jedoch 2-Anforderungen in einem Hintergrundspeicher mit $S_2 = 10,20,40$ ausgelagert werden, bleibt B_1 bis zur Sättigung dieses zusätzlichen Speichers praktisch konstant. In diesem flachen Bereich soll die FG-BG-Strategie deshalb betrieben werden.



Foreground-Background Strategie.
Bild 7.6: Verlustwahrscheinlichkeit B_2 als Funktion des angebotenen Gesamtverkehrs $A=A_1+A_2$.



Foreground-Background Strategie.
Bild 7.5: Verlustwahrscheinlichkeit B_1 als Funktion des angebotenen Gesamtverkehrs $A=A_1+A_2$.

b) Verlustwahrscheinlichkeit für 2-Anforderungen

In Bild 7.6 ist die Verlustwahrscheinlichkeit B_2 für die gleichen Parameter dargestellt. Die Kurven für $S_2 = 20$ sind jedoch zwecks Übersichtlichkeit weggelassen. Aus den Kurven ist ersichtlich, inwiefern durch Hinzufügung von zusätzlichen Speicherplätzen (Hintergrundspeicher), die für 2-Anforderungen reserviert bleiben, die Verlustwahrscheinlichkeit B_2 gesenkt werden kann.

Ferner läßt sich feststellen, daß für einen festen Wert von S_2 die Steigung der Kurven bei ansteigendem Angebot A_1 immer größer werden. Dieses wird durch zwei Effekte verursacht. Zum einen werden 1-Anforderungen bevorzugt abgefertigt und dadurch gibt es einen Stau von 2-Anforderungen. Zum anderen werden die 2-Anforderungen zusätzlich benachteiligt, da mit steigendem Angebot A_1 jeder freiwerdende Platz im Vordergrundspeicher mit höherer Wahrscheinlichkeit von einer 1-Anforderung belegt wird.

7.6.2 Transiente Ergebnisse

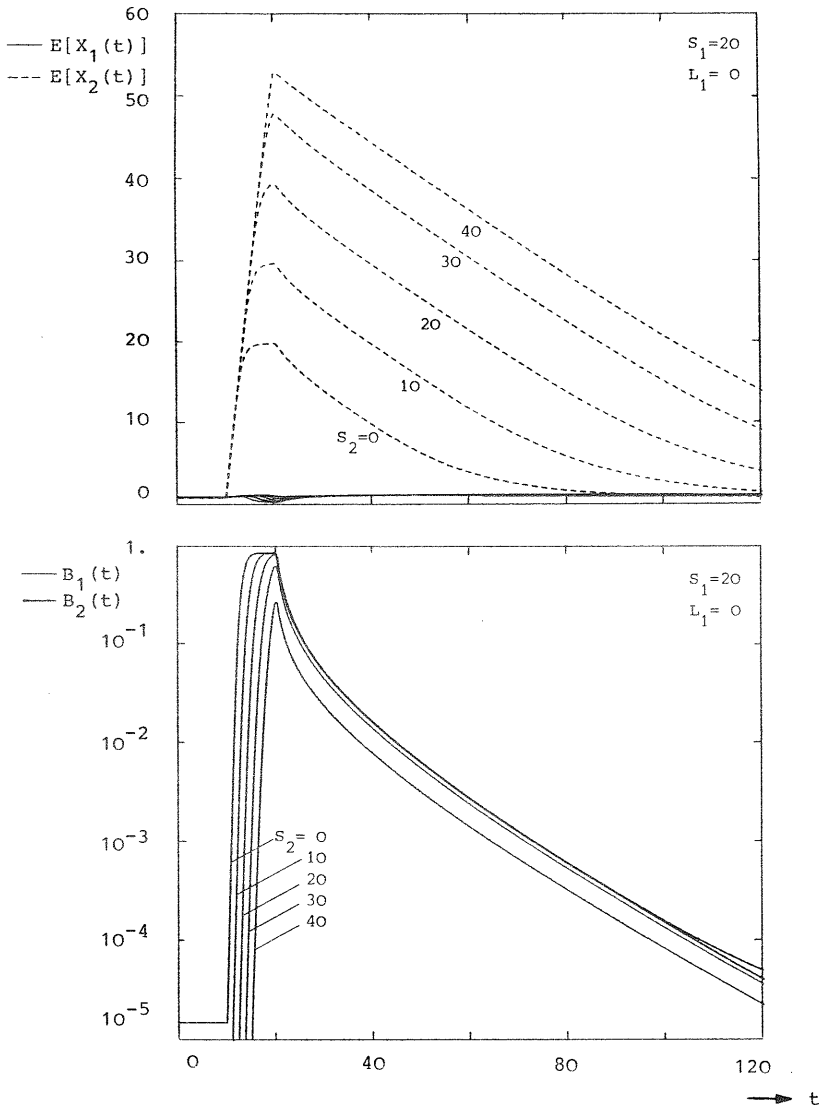
Zur Untersuchung der dynamischen Eigenschaften der FG-BG-Strategie wird im folgenden die Ankunftsrate von einer der beiden Prioritätsklassen rechteckförmig geändert. Betrachtet wird dabei ein Warteschlangensystem mit Vordergrundspeicher $S_1 = 20$ und einem Hintergrundspeicher von jeweils $S_2 = 0,10,20,30$ und 40. Der Überlastimpuls beginnt bei der normierten Zeit $t = 10$ und endet entweder bei $t = 20$ (kurze Überlast) oder bei $t = 40$ (längere Überlast). Ferner gilt für die mittlere Bedienungszeit $h_1 = h_2 = 1$.

7.6.2.1 Kurzer Überlastimpuls der 2. Prioritätsklasse

Betrachtet sei eine stationäre Verkehrssituation mit Ankunftsraten $\lambda_1 = 0.4$ und $\lambda_2(\infty) = 0.2$. Während der Überlastsituation mit Zeitdauer $T = 10$ ändert sich $\lambda_2(t)$ nach einer Rechteckfunktion mit Höchstwert $\lambda_{max} = 6.0$. Ferner werden keine Speicherplätze für 1-Anforderungen reserviert ($L_1 = 0$).

a) Mittlere Systembelastung

Wie Bild 7.7 erwartungsgemäß zeigt, wächst die mittlere Systembelastung $E[X_2(t)]$ durch den rechteckigen Überlastimpuls bis zur Systemsättigung linear an. Die Rate beträgt $(\lambda_{max}-\mu) + \lambda_1 = 5.4$,



Foreground-Background Strategie: Kurzer Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 10.

Bild 7.7: Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 7.8: Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

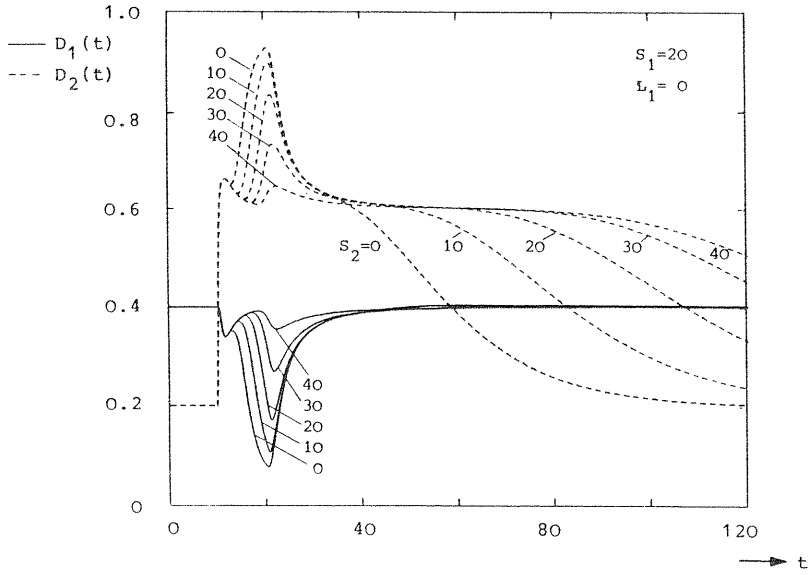
so daß bei genügender Systemkapazität am Ende der Überlastsituation eine mittlere Systembelastung $E[X_2(t)] = 1.75 + 54.0 = 55.75$ erreicht würde. An der Sättigungsgrenze werden die Kurven jeweils abgebogen und schließlich auf die betreffende Speicherkapazität begrenzt. Auch bei einem Hintergrundspeicher $S_2 = 40$ tritt eine geringe Sättigung auf und dadurch steigt die mittlere Systembelastung der 2-Anforderungen nur bis etwa $E[X_2(t)] = 52$ an. Wie in Abschnitt 5.4.3 sinkt die mittlere Systembelastung $E[X_1(t)]$ im Falle eines gesättigten Speichers auf einen sehr niedrigen Wert ab. Die Abnahme ist bei $S_2 = 40$ am geringsten.

b) Verlustwahrscheinlichkeit

In Bild 7.8 sind die Verlustwahrscheinlichkeiten $B_1(t)$ und $B_2(t)$ dargestellt. Da den überlastverursachenden 2-Anforderungen die gesamte Speicherkapazität zur Verfügung steht, liegt aus deren Sicht ein gemeinsamer Speicher der Größe $S = S_1 + S_2$ vor. Die 1-Anforderungen haben nur Zugang zum Vordergrundspeicher, was bedeutet, daß wenn 2-Anforderungen abgewiesen werden, gleichzeitig auch die 1-Anforderungen keinen freien Speicherplatz mehr vorfinden. Deshalb ist im Überlastbereich $B_1(t) = B_2(t)$. Auch diese Kurven zeigen, daß eine kurze Überlast umso besser gemeistert werden kann, je größer die Kapazität S_2 des Hintergrundspeichers gewählt wird. Bei $S_2 = 40$ bleibt die Verlustwahrscheinlichkeit noch unter 0.57, bei $S_2 = 70$ noch unter 10^{-3} [van As (1984a)]. Zu beachten ist auch die lange Systemerholungszeit, wenn die Sättigung erreicht worden ist. Während dieser Zeit ist das System viel empfindlicher für eine neue Überlast.

c) Durchsatz

In Bild 7.9 wird das dynamische Verhalten der Durchsatzkurven betrachtet. Für $S_2 = 0$ wurde der prinzipielle Kurvenverlauf bereits im Abschnitt 5.4.3.c (Kurven 2) diskutiert. Allerdings war dort $\lambda_1 = 0.2$ und dauerte der Überlastimpuls länger, so daß völlige Systemsättigung erreicht wurde. Aus dem vorliegenden Bild läßt sich nun der Einfluß der Hintergrundspeichergröße S_2 entnehmen: je größer S_2, umso weniger wird der Durchsatz der 1. Prioritätsklasse von einem kurzen Überlastimpuls von 2-Anforderungen beeinträchtigt und umso länger arbeitet die Bedienungseinheit mit voller Kapazität, um das aufgenommene Verkehrsvolumen abzufertigen.



Foreground-Background Strategie: Kurzer Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 10.

Bild 7.9: Durchsatz $D_1(t)$ bzw. $D_2(t)$.
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

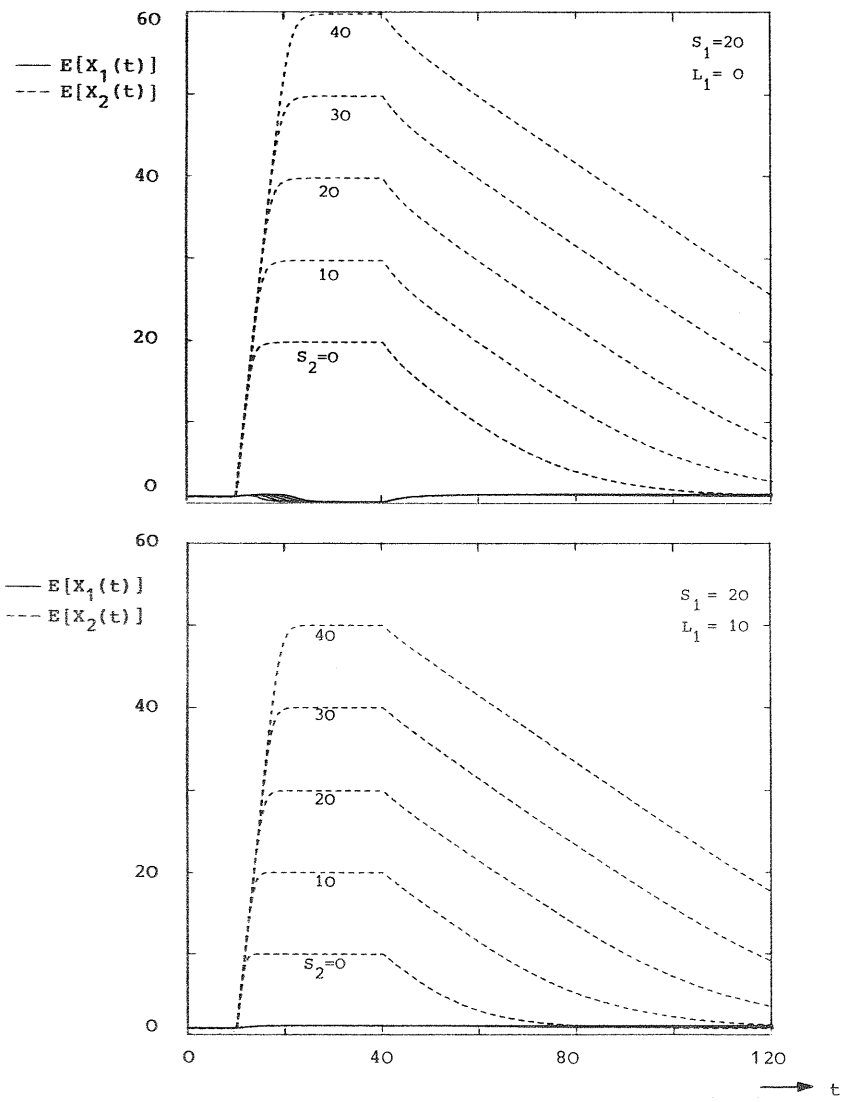
7.6.2.2 Längerer Überlastimpuls der 2. Prioritätsklasse

Betrachtet seien wieder die stationären Ankunftsraten $\lambda_1 = 0.4$ und $\lambda_2(\infty) = 0.2$. Der rechteckige Überlastimpuls ($\lambda_{max} = 6.0$) dauert jetzt aber von $t = 10$ bis $t = 40$.

a) Mittlere Systembelastung

Ausgangspunkt für die Ergebnisdiskussion ist Bild 7.10, wobei durch den lang anhaltenden Überlastimpuls von 2-Anforderungen alle betrachteten Systemkonfigurationen einen gesättigten Zustand erreichen. Der Kurvenanfang deckt sich mit dem in Bild 7.7, wo die Auswirkung eines Überlastimpulses mit gleicher Intensität, aber kürzerer Dauer betrachtet wird. Durch die Speicherverstopfung wird der Verkehrsstrom der 1. Priorität stark behindert, denn jeder freiwerdende Speicherplatz wird im Verhältnis der Ankunftsraten neu belegt. Während der Überlastsituation ist die Wahrscheinlichkeit für eine Neubelegung durch eine 2-Anforderung 15-mal größer. Die Kurven zeigen deshalb einen Rückgang der mittleren Systembelastung $E[X_1(t)]$ auf sehr geringe Werte.

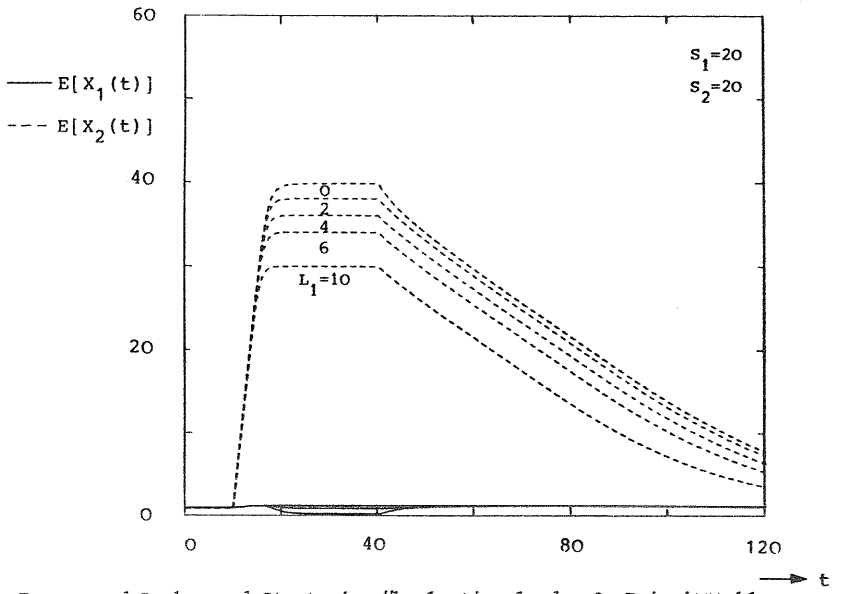
Um diese unerwünschte Tatsache zu beseitigen, wird in Bild 7.11 eine Speicherreservierung $L_1 = 10$ für die 1-Anforderungen vorgenommen. Wie aus den Kurven hervorgeht, erreicht die mittlere Systembelastung $E[X_2(t)]$ jeweils einen Höchstwert von $S_1 + S_2 - L_1$. Insbesondere werden jetzt aber die 1-Anforderungen nur noch geringfügig von den 2-Anforderungen beeinflußt, und zwar nur noch durch eine Restbedienungszeit. Dies macht sich in den Kurven durch eine geringe Erhöhung von $E[X_1(t)]$ während der Überlastsituation bemerkbar. Durch die Anwesenheit der normalen bzw. etwas erhöhten mittleren Systembelastung $E[X_1(t)]$ fehlt der etwas steilere Abfall für $E[X_2(t)]$ direkt beim Verschwinden des Überlastimpulses. Deshalb erreichen diese Kurven ihre stationären Endwerte etwas später als diejenigen ohne Reservierung für 1-Anforderungen. Dabei werden jeweils zwei Kurven mit gleichem Sättigungswert für $E[X_2(t)]$ miteinander verglichen, was durch die spezielle Wahl von $L_1 = 10$ möglich ist.



Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.
Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 7.10: keine Reservierung für die 1. Prioritätsklasse (Dialogpakete), $L_1 = 0$.

Bild 7.11: Reservierung, $L_1 = 10$.
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.



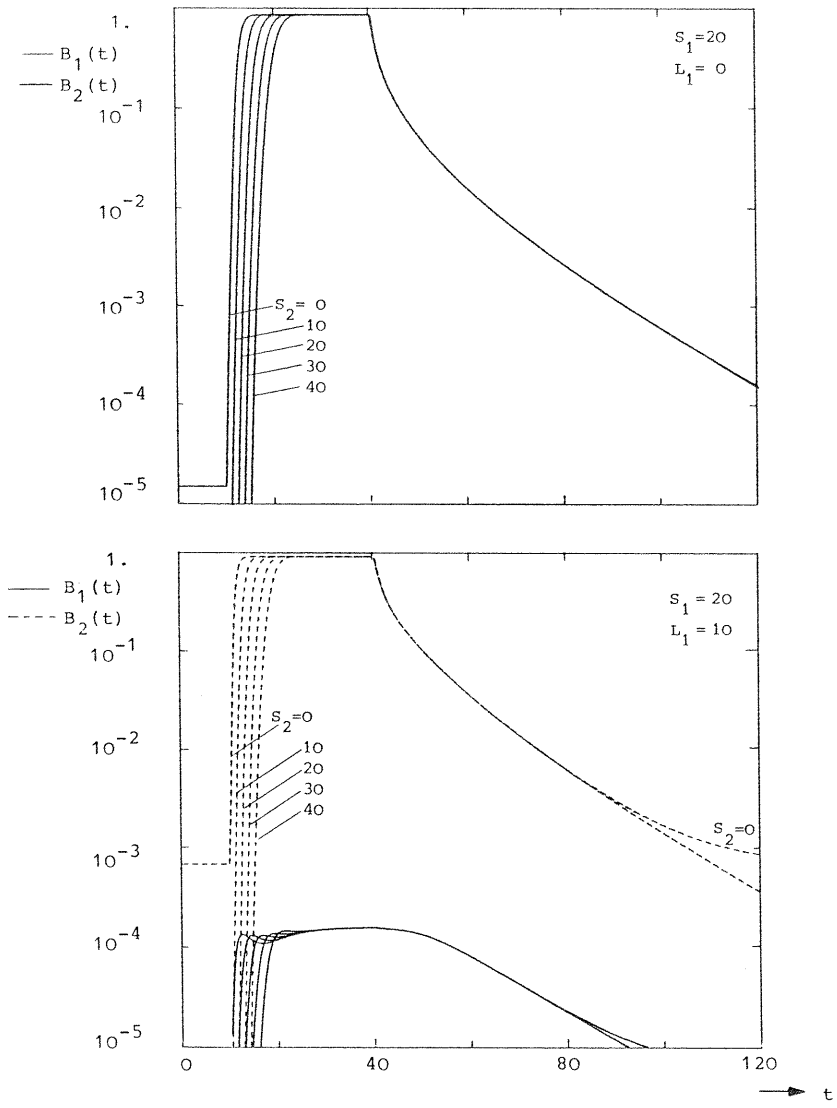
Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.
Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 7.12: Verschiedene Reservierungswerte L_1 für die 1. Prioritätsklasse (Dialogpakete).
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

In Bild 7.12 sind die Kurven für einen Hintergrundspeicher der Größe $S_2 = 20$ und verschiedener Reservierungswerte $L_1 = 0,2,4,6,10$ dargestellt. Da für den stationären Fall die mittlere Systembelastung $E[X_1(\infty)] = 2.5$ beträgt, nimmt bei einer Speicherplatzreservierung $L_1 \geq 4$ die mittlere Systembelastung der 1-Anforderungen während des Überlastimpulses nicht mehr ab.

b) Verlustwahrscheinlichkeit

Als Referenzkurven sind in Bild 7.13 die Verlustwahrscheinlichkeiten $B_1(t)$ bzw. $B_2(t)$ für den längeren Überlastimpuls nochmals dargestellt. Ohne Speicherplatzreservierung ($L_1 = 0$) sind die Verlustwahrscheinlichkeiten für beide Prioritätsklassen identisch.



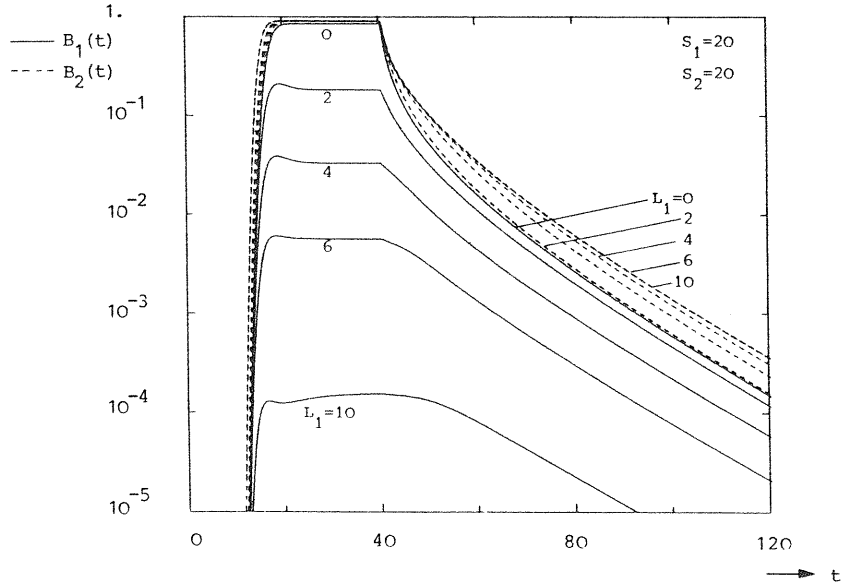
Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.
Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.

Bild 7.13: Keine Reservierung für die 1. Prioritätsklasse (Dialogpakete), $L_1 = 0$.

Bild 7.14: Reservierung, $L_1 = 10$.
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

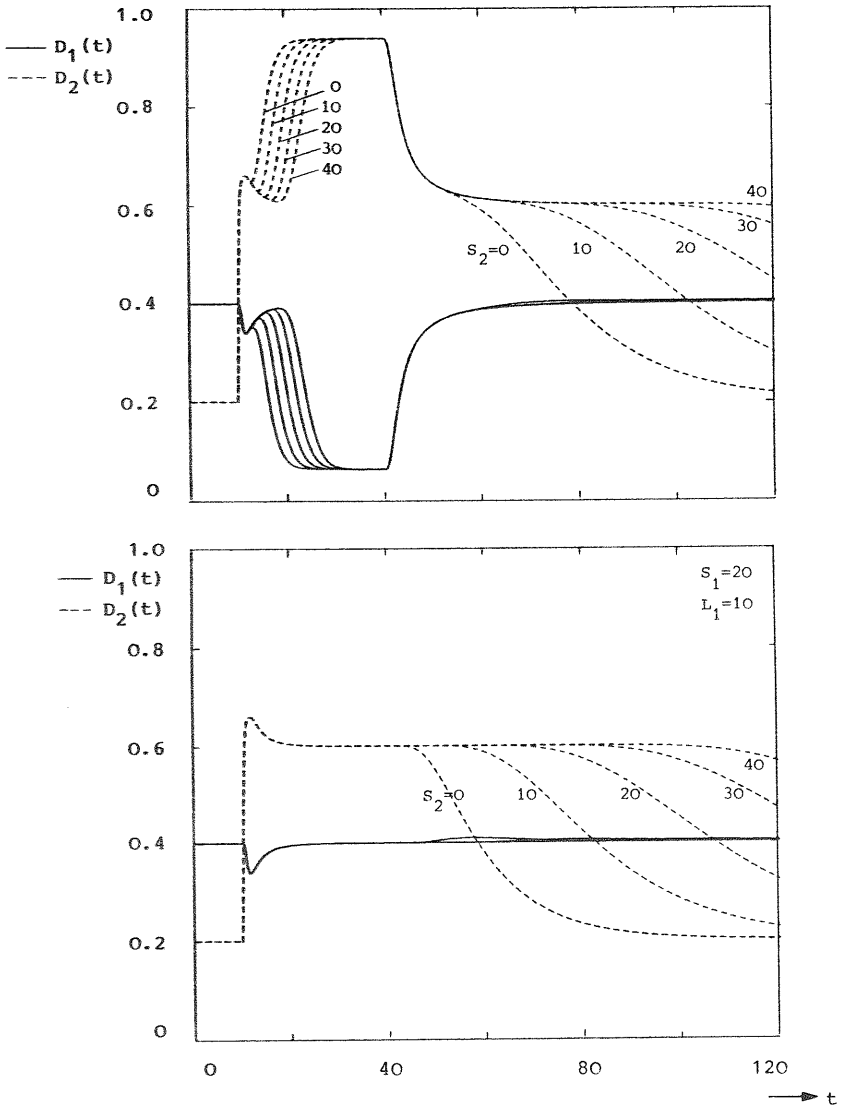
Wird jedoch eine Speicherplatzreservierung für die 1-Anforderungen ($L_1 = 10$) vorgenommen, so kann, wie aus Bild 7.14 hervorgeht, die Verlustwahrscheinlichkeit $B_1(t)$ wesentlich verbessert werden. Die Reservierung reduziert jedoch die den 2-Anforderungen zugängliche Speicherkapazität. Entsprechend früher steigen dann auch die Kurven für $B_2(t)$ an. Aus dem gleichen Grund liegen diese Kurven stets oberhalb der Referenzkurven von Bild 7.13.

Bild 7.15 zeigt, für den Fall $S_2 = 20$, wie die Verlustwahrscheinlichkeit $B_1(t)$ durch eine größere Speicherreservierung für 1-Anforderungen schrittweise verbessert werden kann. Auch die verhältnismäßig geringe Erhöhung der Verlustwahrscheinlichkeit $B_2(t)$ geht aus diesem Bild gut hervor.



Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.
Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.

Bild 7.15: Verschiedene Reservierungswerte L_1 für die 1. Prioritätsklasse (Dialogpakete).
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.



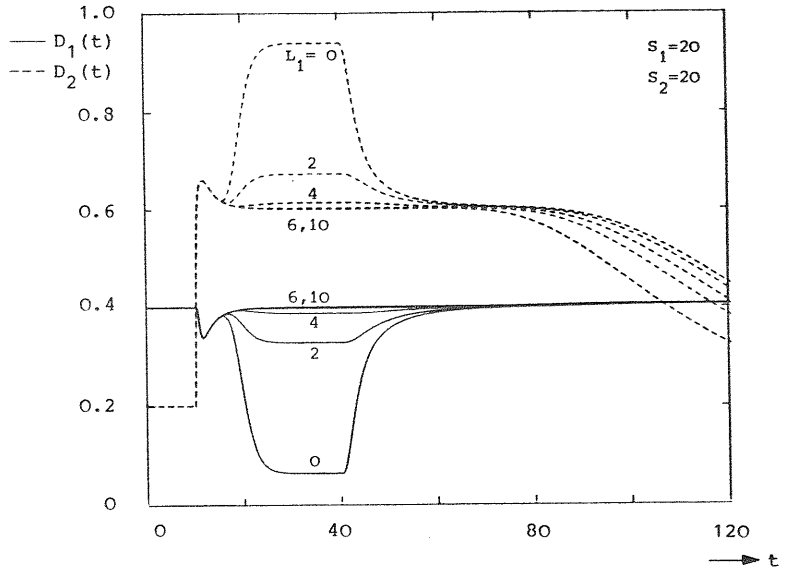
Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.
Durchsatz $D_1(t)$ bzw. $D_2(t)$.

Bild 7.16: Keine Reservierung für die 1. Prioritätsklasse (Dialogpakete), $L_1 = 0$.

Bild 7.17: Reservierung, $L_1 = 10$.
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

c) Durchsatz

In Bild 7.16 ist der zeitliche Verlauf der Durchsatzkurven $D_1(t)$ bzw. $D_2(t)$ für den längeren Überlastimpuls der 2-Anforderungen dargestellt (vgl. auch Bild 7.9 und den dazu gehörenden Text). Wesentlich ist vor allem, daß der Durchsatz der 1-Anforderungen durch eine lang anhaltende Überlast beachtlich absinkt. Eine Vergrößerung des Hintergrundspeichers verzögert lediglich den Zeitpunkt für diese Durchsatzverschlechterung.
Wird jedoch eine Speicherplatzreservierung $L_1 = 10$ für die 1-Anforderungen vorgesehen, so werden die 1-Anforderungen, abgesehen von dem kurzen Durchsatzeinbruch direkt am Anfang des Überlastimpulses, ohne Leistungsabfall (d.h. $D_1(t) = 0.4$) durchgesetzt. Durch diese Maßnahme bleibt der Durchsatz $D_2(t)$ auf den komplementären Maximalwert 0.6 begrenzt. Dies ist in Bild 7.17 wiedergegeben.



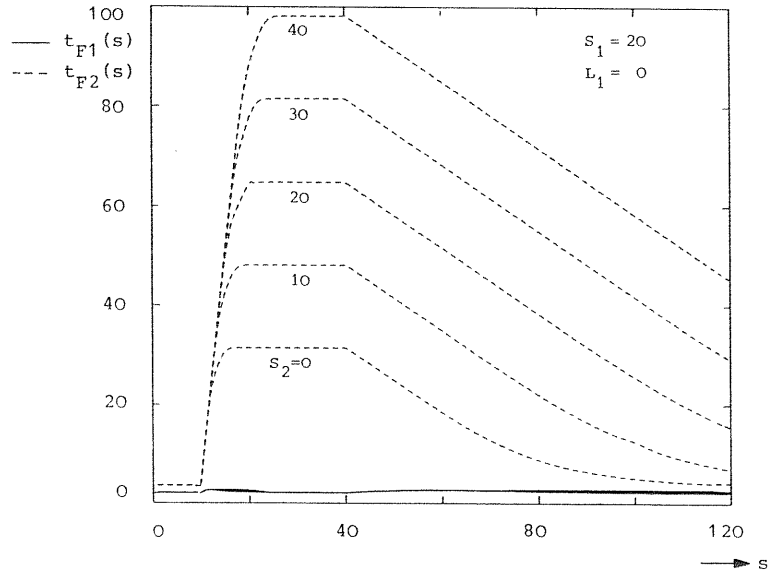
Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.
Durchsatz $D_1(t)$ bzw. $D_2(t)$.

Bild 7.18: Verschiedene Reservierungswerte L_1 für die 1. Prioritätsklasse (Dialogpakete)
Verkehrsparameter: $\lambda_1 = 0,4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

Bild 7.18 zeigt die Auswirkung verschiedener Reservierungswerte L_1 auf den zeitlichen Verlauf der Durchsatzkurven. Für den untersuchten Verkehrsfall ist der Durchsatz $D_1(t)$ bereits mit einer Reservierung von zwei Speicherplätzen für die 1-Anforderungen ($L_1 = 2$) wesentlich unempfindlicher gegen eine Überlastung durch 2-Anforderungen.

d) Mittlere Durchlaufzeit

Bild 7.19 zeigt die mittlere Durchlaufzeit $t_{F1}(s)$ bzw. $t_{F2}(s)$ einer zum Zeitpunkt s eintreffenden, akzeptierten 1- bzw. 2-Anforderung. Da ein Überlastimpuls der 2. Prioritätsklasse vorliegt, wird $t_{F1}(s)$ nur durch die erhöhte Wahrscheinlichkeit einer Restbedienung einer 2-Anforderung beeinflußt. Somit weist die mittlere Durchlaufzeit $t_{F1}(s)$ etwas höhere Werte auf. Bedingt durch die zeitunabhängige Rate λ_1 ist der zeitliche Verlauf der mittleren Durchlaufzeit $t_{F2}(s)$ durch den angetroffenen Systemzustand bestimmt. Der Durchlaufprozeß der betrachteten 2-Anforderung verläuft danach wie im stationären Fall. Die Kurven für $t_{F2}(s)$ zeigen deshalb einen ähnlichen Verlauf wie die entsprechenden Kurven für die mittlere Systembelastung $E[X_2(t)]$ in Bild 7.10.



Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Stapelpakete), Impulsdauer T = 30.

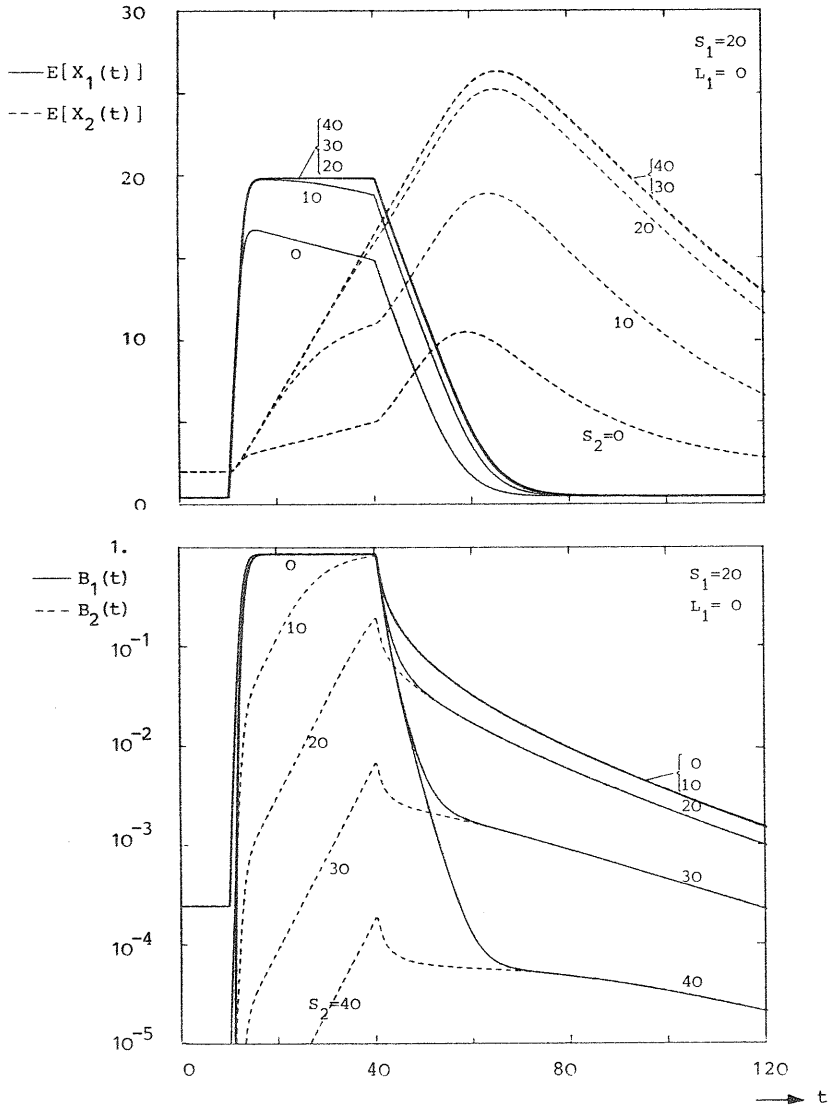
Bild 7.19: Mittlere Durchlaufzeit $t_{F1}(s)$ bzw. $t_{F2}(s)$ als Funktion vom Ankunftszeitpunkt s.
Verkehrsparameter: $\lambda_1 = 0.4$, $\lambda_2(\infty) = 0.2$, $\lambda_{2MAX} = 6.0$, $h_1 = h_2 = 1$.

7.6.2.3 Längerer Überlastimpuls der 1. Prioritätsklasse

Als nächstes wird die Systemreaktion auf einen Überlastimpuls von 1-Anforderungen untersucht. Die stationären Ankunftsraten zum Beobachtungsbeginn seien $\lambda_1 = 0.2$ und $\lambda_2 = 0.5$. Die konstante Rate ist diesmal etwas höher gewählt, um die Auswirkungen etwas deutlicher hervorheben zu können. Der rechteckige Überlastimpuls mit $\lambda_{max} = 6.0$ fängt wieder bei $t = 10$ an und endet bei $t = 40$.

a) Mittlere Systembelastung

In Bild 7.20 erkennt man zuerst den Speicherbelegungsaustausch der beiden Prioritätsklassen, wenn nur der gemeinsame Vordergrundspeicher zur Verfügung steht ($S_2 = 0$, vgl. Abschnitt 5.4.3b). Wird ein Hintergrundspeicher verwendet, so kann bei genügender Kapazität (im vorliegenden Fall $S_2 = 30$ bzw. $S_2 = 40$) der komplette Vordergrundspeicher durch Auslagerung der 2-Anforderungen für die 1. Prioritätsklasse freigemacht werden. Darüberhinaus können die während der Überlastsituation eintreffenden 2-Anforderungen vom Hintergrundspeicher aufgenommen werden. Bei einem zu kleinen Hintergrundspeicher ($S_2 = 10$) oder bei einer noch länger anhaltenden Überlast behindern sich die beiden Prioritätsklassen wieder gegenseitig.



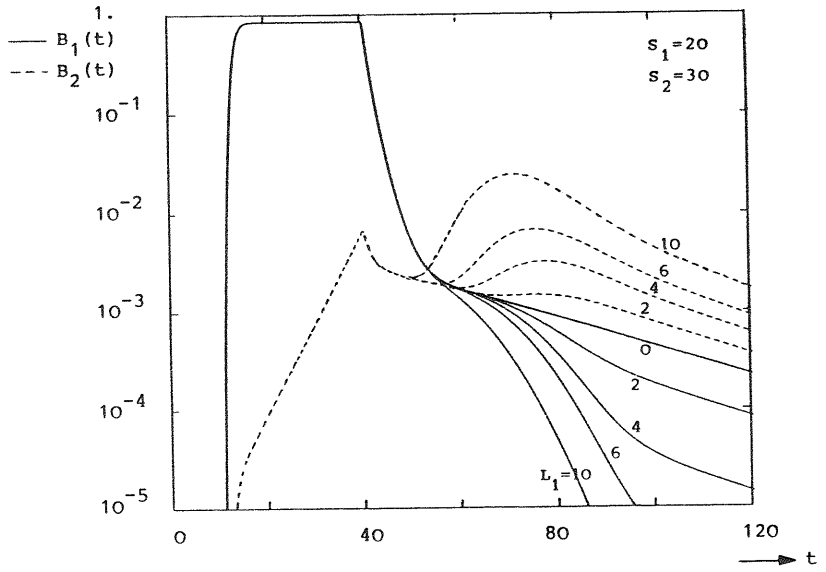
Foreground-Background Strategie: Überlastimpuls der 1. Prioritätsklasse (Dialogpakete), Impulsdauer T = 30.

Bild 7.20: Mittlere Systembelastung $E[X_1(t)]$ bzw. $E[X_2(t)]$.

Bild 7.21: Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.5$, $h_1 = h_2 = 1$.

b) Verlustwahrscheinlichkeit

Bild 7.21 zeigt den dynamischen Verlauf der beiden Verlustwahrscheinlichkeiten $B_1(t)$ und $B_2(t)$. Durch die zum Zeitpunkt $t = 10$ schlagartig erhöhte Ankunftsrate $\lambda_1(10) = 6.0$ läuft der Vordergrundspeicher schnell voll und die bereits anwesenden 2-Anforderungen werden, falls ein Hintergrundspeicher vorhanden ist, suksessive ausgelagert. Sobald der Vordergrundspeicher voll belegt ist und eine Auslagerung nicht mehr möglich ist, werden die 1-Anforderungen abgewiesen (maximaler Wert von $B_1(t)$).



Foreground-Background Strategie: Überlastimpuls der 1. Prioritätsklasse (Dialogpakete), Impulsdauer T = 30.

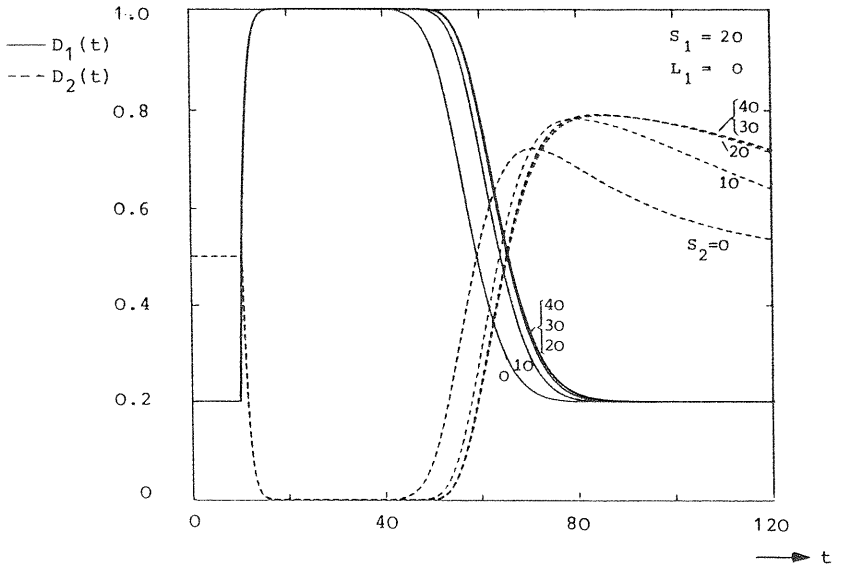
Bild 7.22: Verlustwahrscheinlichkeit $B_1(t)$ bzw. $B_2(t)$ bei Reservierung L_1 für die 1. Prioritätsklasse.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.5$, $h_1 = h_2 = 1$.

Durch die restriktive Speicherplatzbelegung für Anforderungen der 1. Prioritätsklasse wird die Verlustwahrscheinlichkeit $B_2(t)$ nur teilweise beeinflußt. Dies drückt sich in den Kurven durch den Übergang von einer anfangs sehr steilen in eine lineare Zunahme von $B_2(t)$ aus. In Sättigungsnähe ($S_2 = 10$) werden die Kurven abgeflacht. Bei großer Speicherkapazität ($S_2 = 40$) steigt $B_2(t)$ nur gering an. Nach Verschwinden der Überlast nimmt $B_1(t)$ umso schneller ab, je größer S_2 ist. Auch $B_2(t)$ fällt im ersten Augenblick steil ab. Danach sind die beiden Kurven im betrachteten Zeitintervall deckungsgleich; denn während des Abbaus von Anforderungen der 1. Prioritätsklassen haben sich weitere 2-Anforderungen rückgestaut, so daß der zusammenfallende Kurvenverlauf die Wahrscheinlichkeit für einen vollbelegten Gesamtspeicher $S = S_1+S_2$ angibt.

An Hand von Bild 7.22 wird die Auswirkung einer Speicherplatzreservierung ($L_1 = 0,2,4,5,10$) für 1-Anforderungen diskutiert. Diese Reservierung war notwendig, um die Aufnahme von Anforderungen der 1. Prioritätsklasse bei einer Überlastung durch 2-Anforderungen garantieren zu können (vgl. Abschnitt 7.6.2.2). Betrachtet wird insbesondere die Hintergrundspeicherkapazität $S_2 = 30$. Wie aus den Kurven ersichtlich ist, fällt die Verlustwahrscheinlichkeit $B_1(t)$ am Ende des Überlastimpulses mit zunehmendem Wert L_1 schneller ab, währenddessen die Kurven für $B_2(t)$ nochmals ansteigen. Der Grund für dieses Verhalten ist, daß durch die Reservierung einerseits weniger 1-Anforderungen abgewiesen werden, andererseits aber die 2-Anforderungen jetzt über einen kleineren gemeinsamen Speicher verfügen. Die entsprechende Kurve für die mittlere Systembelastung $E[X_2(t)]$ liegt deshalb ab $t = 70$ auch geringfügig unterhalb der korrespondierenden Kurve von Bild 7.20. Weitere numerische Ergebnisse bezüglich dieses Effektes sind in [van As (1984a)] angegeben.

c) Durchsatz

In Bild 7.23 sind die Kurven für den Durchsatz $D_1(t)$ bzw. $D_2(t)$ dargestellt. Als Referenz wird das System ohne Hintergrundspeicher ($S_2 = 0$, vgl. Abschnitt 5.4.3c) betrachtet. Wie aus dem Verlauf von $D_1(t)$ ersichtlich ist, hält der hohe Durchsatz von 1-Anforderungen umso länger an, je weniger 2-Anforderungen im Vordergrundspeicher vorhanden sind. Die anschließende Erhöhung von $D_2(t)$ ist abhängig vom zurückgestellten Verkehrsvolumen.



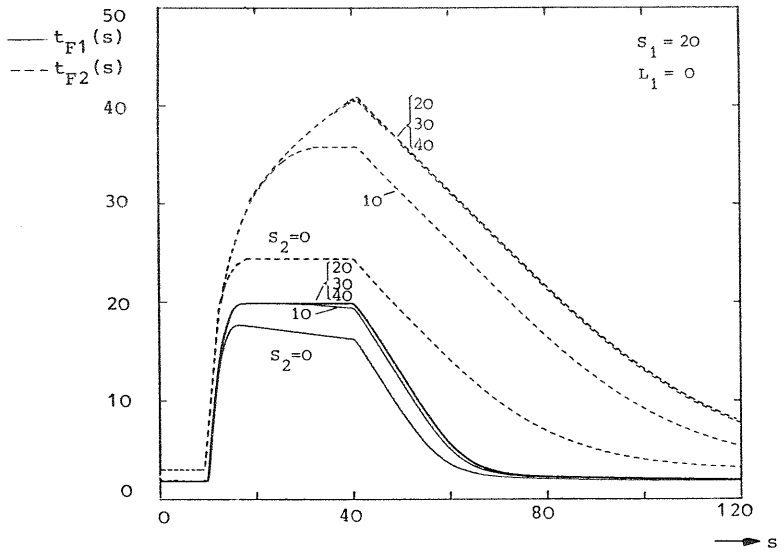
Foreground-Background Strategie: Überlastimpuls der 2. Prioritätsklasse (Dialogpakete), Impulsdauer T = 30.

Bild 7.23: Durchsatz $D_1(t)$ bzw. $D_2(t)$.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.5$, $h_1 = h_2 = 1$.

d) Mittlere Durchlaufzeit

In Bild 7.24 ist die mittlere Durchlaufzeit $t_{F1}(s)$ bzw. $t_{F2}(s)$ einer zum Zeitpunkt s eintreffenden, akzeptierten 1- bzw. 2-Anforderung aufgezeichnet. Durch die Systemsättigung bei den Systemparametern $S_2 = 0$ und $S_2 = 10$ nimmt $t_{F1}(s)$ während der Überlast ab und bleibt $t_{F2}(s)$ während dieser Zeit konstant.

Die mittlere Durchlaufzeit $t_{F1}(s)$ wird geringer, weil durch zurückgestellte 2-Anforderungen immer weniger 1-Anforderungen im System anwesend sein können. Für eintreffende 2-Anforderungen ist nur die Gesamtanzahl von vorhandenen Anforderungen von Bedeutung. Im gesättigten Zustand ist ihre Durchlaufzeit somit konstant. Als Folge der schnellen Zunahme der mittleren Systembelastung $E[X_1(t)]$ nimmt $t_{F2}(s)$ am Anfang schnell zu. Der weitere Verlauf der Kurven für $t_{F2}(s)$ wird von der Größe des Hintergrundspeichers und von der Ankunftsrate der 2-Anforderungen bestimmt. Für die betrachteten Verkehrsparameter sind die Kurven ab einer Hintergrundspeichergröße $S_2 = 20$ deckungsgleich. Eine Sättigung kommt dann nicht zustande.



Foreground-Background Strategie: Überlastimpuls der 1. Prioritätsklasse (Dialogpakete), Impulsdauer T = 30.

Bild 7.24: Mittlere Durchlaufzeit $t_{F1}(s)$ bzw. $t_{F2}(s)$ als Funktion vom Ankunftszeitpunkt s.
Verkehrsparameter: $\lambda_1(\infty) = 0.2$, $\lambda_{1MAX} = 6.0$, $\lambda_2 = 0.5$, $h_1 = h_2 = 1$.

7.7 Schlußfolgerung

In einem Paketvermittlungsnetz mit Benutzerprioritäten (z.B. Dialog- und Stapelbetrieb) ist die Foreground-Background Überlastabwehrstrategie eine geeignete und schnell wirkende Maßnahme, um kurze Überlastsituationen im Netzinnern aufzufangen. Bei länger anhaltender Überlast sind aber zusätzliche Maßregeln notwendig.

Ferner ermöglicht sie eine verbesserte Speicheroptimierung. Wird zum Beispiel ein modularer Netzknoten betrachtet, so kann bei einer fest dimensionierten Speichergröße S_1 einer Vermittlungseinheit ihre Speicherungsmöglichkeit trotzdem an die zu erwartende Verkehrssituation (Verkehrscharakteristik, maximale Überlastspitzen) angepaßt werden. Die Anpassung geschieht mit zwei Softwaregrößen: der zusätzliche Speicherbereich S_2 in einem billigeren und für alle Vermittlungseinheiten gemeinsamen Hintergrundspeicher für die zurückgestellten Pakete niedriger Priorität und der reservierte Speicherbereich L_1 im Vermittlungsspeicher selbst für die Pakete höherer Priorität. Für eine vergleichbare Verkehrsleistung ohne Hintergrundspeicher muß der Speicher jeder Vermittlungseinheit überdimensioniert werden. Bei normalen Verkehrsverhältnissen wird dann der Speicher sehr schlecht ausgenutzt.

8. ÜBERLASTABWEHR DURCH ZWEIPUNKT-REGELUNG

In den beiden vorangegangenen Kapiteln wurden Überlastabwehrstrategien behandelt, die lokal wirken. Dies bedeutet, erkennende und ausführende Funktionen sind in den betrachteten Netzknoten angesiedelt. Im folgenden werden global wirkende Strategien behandelt. Hier sind die notwendigen Zustandsinformationen bzw. die auszulösenden Maßnahmen durch Meldungen von bzw. zu anderen Netzknoten bereitzustellen.

8.1 Allgemeines

In diesem Kapitel wird eine Überlastabwehrstrategie untersucht, die auf dem Mechanismus einer Zweipunkt-Regelung basiert. Bei dieser Strategie, die als Ergänzung zu den viel schnelleren lokalen Maßnahmen vorgesehen ist, wird versucht, das Entstehen oder die Ausbreitung einer Überlastsituation mit Hilfe von Meldungen an die betreffenden Verkehrsquellen zu verhindern.

Das Verkehrsangebot wird dazu zustandsabhängig gedrosselt, so daß bei richtiger Wahl der relevanten Parameter ein maximaler Paketdurchsatz erreicht werden kann. Dies gelingt, wenn einerseits die Blindlast, die durch Paketwiederholungen bei einem Speicherüberlauf entsteht, gering gehalten werden kann, andererseits aber verhindert wird, daß der Verkehrsfluß durch zu starke Drosselung unnötig gebremst wird.

Als Entscheidungskriterium für das Aussenden von Meldungen dient der momentane Belegungszustand in den Warteschlangen für die abgehenden Richtungen oder auch die Belegung im gesamten Paketspeicher. Sofern die Verkehrsquellen dem Netzknoten bekannt sind (z.B. bei virtuellen Verbindungen), können die Abwehrmaßnahmen sich in erster Linie auf Verkehrsströme der überlasteten Richtung beschränken. Im weiteren wird der Belegungszustand eines Paketspeichers zur Erkennung einer bevorstehenden Überlastsituation herangezogen.

Der Mechanismus einer Zweipunkt-Regelung ist im Bild 8.1 an einem Paketspeicher, der S Pakete aufnehmen kann, veranschaulicht. Bei dieser Regelung, die bei einer Überlast in Aktion tritt, wird das Verkehrsangebot mit der Ankunftsrate λ_1

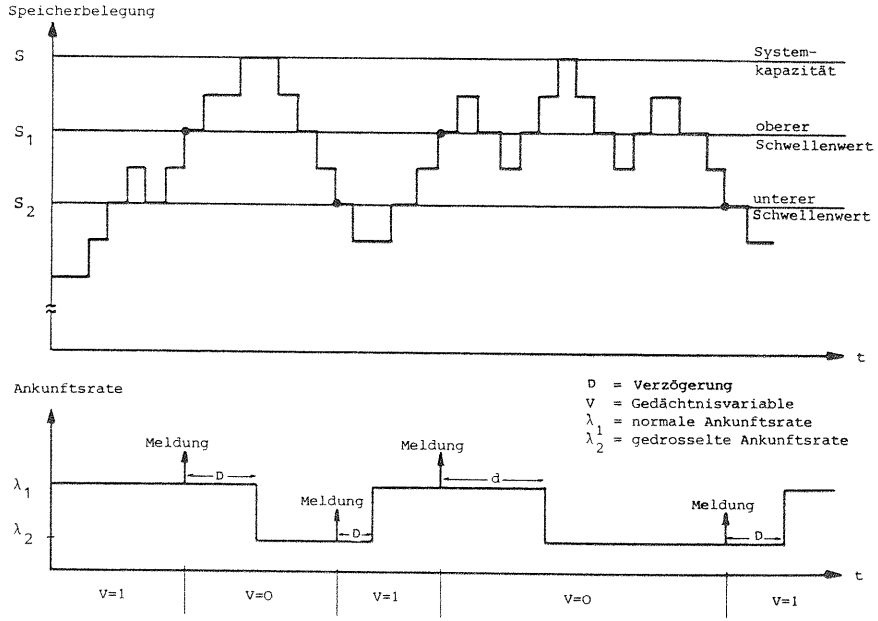


Bild 8.1: Der Mechanismus der Zweipunkt-Regelung mit einer stochastisch verzögerten Wirkung.

(normaler Betrieb) auf die Rate λ_2 gedrosselt. Somit kann der Netzknoten sich wieder von der Überlastsituation erholen, so daß danach die ursprüngliche Ankunftsrate λ_1 erneut zugelassen werden kann.

Für die Regelung sind drei Software-Größen notwendig:

- eine Variable V, die den zuletzt gesendeten Meldungstyp festhält, um zu verhindern, daß zwei identische Meldungen infolge statistischer Schwankungen nacheinander ausgesendet werden (V = 1: normaler Betrieb, V = 0: gedrosselter Betrieb),
- ein oberer Schwellenwert S_1, der das Aussenden einer Meldung zur Drosselung einleitet, wenn die Speicherbelegung den Wert S_1 erreicht und die Variable V einen normalen Betrieb anzeigt,
- ein unterer Schwellenwert S_2, der das Aussenden einer Meldung zur Wiederherstellung des normalen Betriebes einleitet, wenn die Speicherbelegung wieder auf den Wert S_2 abgesunken ist und die Variable V einen gedrosselten Betrieb anzeigt.

Durch die geographische Entfernung zwischen Netzknoten und Verkehrsquellen hat die Regelung eine stochastische Verzögerung. Die Verzögerungszeit bis zum Einsetzen der Regelung setzt sich zusammen aus:

- der Zeit, die vergeht bis die Meldungen bei den Verkehrsquellen eingetroffen sind. Durch die bevorzugte Behandlung der Netzsteuerungspakete, die diese Meldungen enthalten, ist dieser Anteil vergleichsweise gering.
- der Zeit, die verstreicht bis der Einfluß des Verkehrsvolumens, das bereits vor der Änderung der Ankunftsrate im Netz vorhanden war, sich nicht mehr auswirkt, oder bis der Verkehr von der gedrosselten Rate auf die normale Rate angestiegen ist. Erst nach dieser Verzögerungszeit gilt die neue Ankunftsrate.

8.2 Modellbeschreibung

Das Verkehrsmodell nach Bild 8.2 wird beschrieben durch:

- ein Warteschlangensystem mit einer Bedienungseinheit, das Maximal S Anforderungen aufnehmen kann (inklusive der bedienten Anforderung),
- einen Schalter, der entweder die Stellungen EIN/AUS gemäß Bild 8.2a besitzt, oder zwischen einem normalen Ankunftsprozeß (Rate λ_1) und einem gedrosselten Ankunftsprozeß (Rate λ_2) gemäß Bild 8.2b hin- und herschaltet,
- eine stochastische Verzögerung zwischen dem Senden einer Meldung für den Schalter und der endgültigen Auswirkung auf den Verkehrsfluß.

Das Warteschlangensystem stellt den gesamten Paketspeicher in einem Netzknoten dar, wobei die Pakete mit einer mittleren Bedienungsrate μ übermittelt werden. Das gleiche Modell gilt sinngemäß auch, wenn die Untersuchung sich nur auf eine bestimmte Richtung bezieht und somit ein einziger Übertragungskanal mit seiner Warteschlange modelliert wird. Mit Hilfe des Schalters und der stochastischen Verzögerung ist es möglich den Ankunftsprozeß des betrachteten Warteschlangensystems so zu modellieren, daß der Netzeinfluß in grober Form berücksichtigt werden kann.

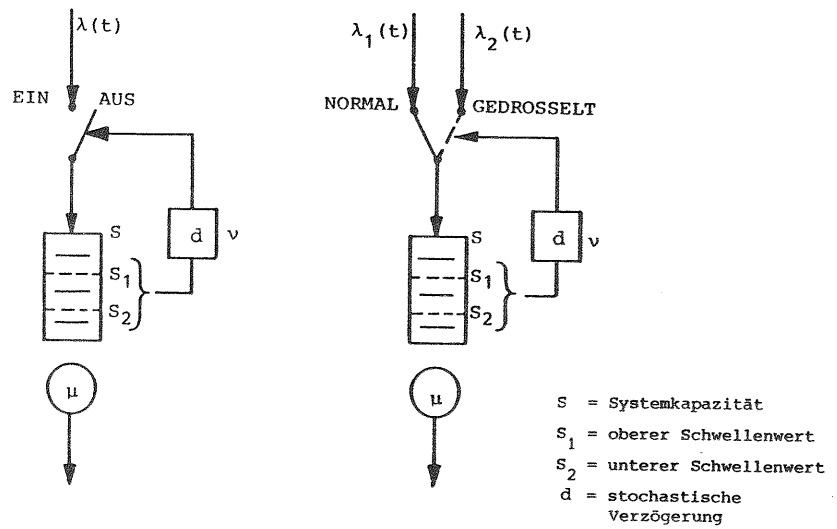


Bild 8.2: Verkehrsmodell: Zweipunkt-Regelung mit einer stochastisch verzögerten Wirkung.
a) Intermittierende Regelung
b) Wechselregelung

Die Zwischenankunftszeiten, die Bedienungszeiten und die Verzögerungszeiten gehorchen jeweils einer negativ exponentiellen Verteilung (M). Die mittlere Bedienungszeit sei $h = 1/\mu$, die mittlere Verzögerungszeit sei $d = 1/\nu$. Anforderungen, die entweder die offene Schalterstellung AUS vorfinden oder im Warteschlangensystem keinen freien Platz antreffen, gehen verloren und haben keinen weiteren Einfluß auf das Systemgeschehen.

8.3 Modellanalyse

Der Zustandsprozeß läßt sich vollständig beschreiben mit einer Zufallsvariable X und zwei Hilfsvariablen Y und Z:

X Anzahl der Anforderungen im System, $X = 0,\ldots,S$

Y Schalterstellung
$Y = 1$: EIN bzw. normaler Betrieb
$Y = 0$: AUS bzw. gedrosselter Betrieb

Z Anzahl der Meldungen die unterwegs sind, $Z = 0,1,\ldots$

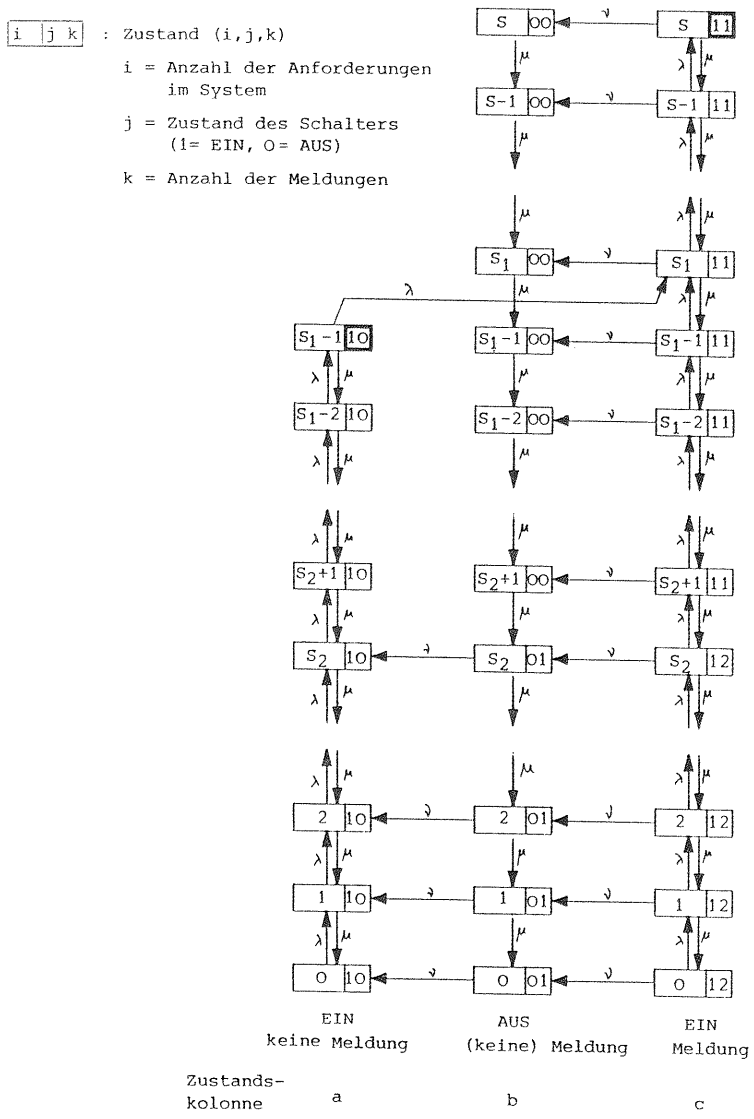


Bild 8.3: Zustandsdiagramm (Intermittierende Regelung, einfachste Form).

Mit der entsprechenden Bezeichnung (i,j,k) für die Zustände kann dieser Prozeß in Form eines Zustandsdiagramms dargestellt werden, das unter den vorausgesetzten Markoff-Bedingungen rekursiv, iterativ oder transient behandelt werden kann. In den nächsten Abschnitten werden drei verschiedene Zustandsräume betrachtet, die sich nach Art der Regelung oder nach Art der Untersuchungen unterscheiden.

8.3.1 Die intermittierende Regelung in ihrer einfachsten Form

a) Zustandsdiagramm

In Bild 8.3 ist die Struktur des Zustandsdiagramms für das Verkehrsmodell mit einer intermittierenden Regelung nach Bild 8.2a dargestellt. Diese Darstellung basiert auf der Annahme, daß neben der Meldung zum Ausschalten höchstens noch eine Meldung zum Wiedereinschalten unterwegs sein darf.

Zur Erläuterung dieses Zustandsdiagramms wird von einem leeren System ausgegangen, das durch den Belegungszustand O mit Zusatzkennung (1,O) charakterisiert wird. Dieses bedeutet, daß der Schalter auf EIN steht und keine Meldungen unterwegs sind.

Vorerst werden die einzelnen Zustände der linken Kolonne in zufälliger Weise durchlaufen. Erreicht aber die Systembelastung den oberen Schwellenwert S_1, so wird die Zustandskolonne gewechselt, und der Zustandsprozeß setzt sich fort über die Zustände mit der Kennung (1,1) die bedeutet, daß vorläufig die Schalterstellung zwar noch im Zustand EIN bleibt, aber eine Meldung zur Betätigung des Schalters unterwegs ist.

Sinkt die Systembelastung während dieser Zeit derart ab, daß der untere Schwellenwert S_2 erreicht wird, so wird das Aussenden einer zweiten Meldung - jetzt zur Wiedereinschaltung des Ankunftsprozesses - veranlaßt. Die diesbezüglichen Zustände haben die Kennung (1,2). Gemäß Annahme ist dies die Höchstzahl von Meldungen, die gleichzeitig unterwegs sein dürfen.

Aus sämtlichen Zuständen dieser Kolonne kann der nächste Kolonnenwechsel mit der Rate ν erfolgen. Abhängig davon, ob dieser Übergang oberhalb oder unterhalb des unteren Schwellenwertes S_2

stattfindet, ist die neue Kennung entweder (0,0) oder (0,1). Die Kennung (0,0) bedeutet, daß die Meldung eingetroffen ist und der Schalter betätigt wurde (Stellung AUS). Der Wechsel zu einem Zustand mit der Kennung (0,1) zeigt an, daß die Meldung zum Ausschalten eingetroffen ist und die Meldung zum Wiedereinschalten bereits unterwegs ist. In der jetzt erreichten Zustandskolonne kann sich das System nur entleeren. Über Zustände mit Kennung (0,1) kann schließlich der Ankunftsprozeß wieder eingeschaltet werden. Dadurch wird die ursprüngliche Zustandskolonne wieder erreicht.

b) Rekursive Berechnung der stationären Zustandswahrscheinlichkeiten

Die Struktur des Zustandsdiagramms für dieses Modell ist gut geeignet für eine rekursive Lösungstechnik (Vgl. Abschnitt 4.4.2). Die stationären Zustandswahrscheinlichkeiten werden deshalb auf diese Weise berechnet. In Übereinstimmung mit der Bezeichnungsweise im Zustandsdiagramm werden die stationären Zustandswahrscheinlichkeiten wie folgt dargestellt:

$$p_i^{jk} = P\{X=i,\ Y=j,\ Z=k\} \quad , \quad i=0,\ldots,S \quad , \quad j=0,1 \quad , \quad k=0,1,2 \ . \tag{8.1}$$

Aufgrund der statistischen Gleichgewichtsbedingung in einem stationären Markoff-Zustandsraum, kann ein System von rekursiven Gleichungen für die stationären Zustandswahrscheinlichkeiten aufgestellt werden. Dabei wird jeweils ein Teil der bereits berechneten Zustände als Makrozustand zusammengefaßt. Da die Zusammenfassung der Zustände im Berechnungsalgorithmus für die Zustandswahrscheinlichkeiten und auch bei der anschließenden Berechnung der charakteristischen Verkehrsgrößen kolonnenweise geschieht, wird für diesen Zweck eine kolonnenorientierte Schreibweise eingeführt:

$$p_i^{jk} = p_i^{\alpha} \qquad \text{mit} \qquad \alpha = \begin{cases} \text{a: Kennung} & (1,0) \\ \text{b: Kennungen} & (0,0)/(0,1) \\ \text{c: Kennungen} & (1,1)/(1,2) \end{cases} \tag{8.2}$$

Unter Berücksichtigung der Berechnungsreihenfolge lassen sich somit die folgenden rekursiven Gleichungen für die Zustandswahrscheinlichkeiten ableiten.

Für die Zustandswahrscheinlichkeiten der Kolonne c gilt:

$$\begin{aligned} p_{i-1}^{c} &= \frac{\mu}{\lambda} \cdot p_i^{c} + \frac{\nu}{\lambda} \cdot \sum_{n=i}^{S} p_n^{c} && , \ i = S,\ldots,S_1+1 \\ p_{i-1}^{c} &= \frac{\mu}{\lambda} \cdot p_i^{c} + \frac{\nu}{\lambda} \cdot \sum_{n=i}^{S} p_n^{c} - p_{S_1-1}^{a} && , \ i = S_1,\ldots,1 \ . \end{aligned} \tag{8.3}$$

Entsprechend gilt für Zustandskolonne b:

$$\begin{aligned} p_S^{b} &= \frac{\nu}{\mu} \cdot p_S^{c} \\ p_i^{b} &= \frac{\nu}{\mu} \cdot p_i^{c} + p_{i+1}^{b} && , \ i = S-1,\ldots,S_2+1 \ , \\ p_i^{b} &= \frac{\nu}{\nu+\mu} \cdot p_i^{c} + \frac{\mu}{\nu+\mu} \cdot p_{i+1}^{b} && , \ i = S_2,\ldots,1 \\ p_0^{b} &= p_0^{c} + \frac{\mu}{\nu} \cdot p_1^{b} \end{aligned} \tag{8.4}$$

und schließlich bekommt man für Zustandskolonne a:

$$\begin{aligned} p_{i-1}^{a} &= \frac{\mu}{\lambda} \cdot p_i^{a} + p_{S_1-1}^{a} && , \ i = S_1-1,\ \ldots,S_2+1 \\ p_{i-1}^{a} &= \frac{\mu}{\lambda} \cdot p_i^{a} + p_{S_1-1}^{a} - \frac{\nu}{\lambda} \cdot \sum_{n=i}^{S_2-i+1} p_n^{b} && , \ i = S_2,\ldots,1 \ . \end{aligned} \tag{8.5}$$

Wie aus den Gleichungen bzw. aus dem Zustandsdiagramm entnommen werden kann, läßt sich dieses rekursive Gleichungssystem berechnen, wenn zwei Zustandswahrscheinlichkeiten als Startpunkte für die Rekursion a-priori festgelegt werden:

$$p_X^1 = p_S^{c} \qquad \text{bzw.} \qquad p_X^2 = p_{S_1-1}^{a} \ . \tag{8.6}$$

Diese Zustände sind im Bild 8.3 speziell gekennzeichnet. Entsprechend dem erläuterten Verfahren in Abschnitt 4.4.2 wird das System von rekursiven Gleichungen (8.3)-(8.5) einmal mit dem Startvektor $(p_X^1,\ p_X^2) = (V,0)$ und einmal mit $(p_X^1,\ p_X^2) = (0,V)$ durchgerechnet. Im allgemeinen wird dabei $V = 1$ gewählt.

Bezeichnet man die korrespondierenden Koeffizienten mit A_i^{jk} bzw. B_i^{jk}, so können alle Zustandswahrscheinlichkeiten als deren Linearkombination ausgedrückt werden:

$$p_i^{jk} = A_i^{jk} \cdot p_X^1 + B_i^{jk} \cdot p_X^2 \ . \tag{8.7a}$$

Oder in der Kolonnenschreibweise:

$$P_i^{\alpha} = A_i^{\alpha} \cdot P_X^1 + B_i^{\alpha} \cdot P_X^2 \quad , \alpha = \text{Kolonne a,b,c} \quad . \tag{8.7b}$$

Hierbei sind die a-priori-Zustandswahrscheinlichkeiten P_X^1 bzw. P_X^2 mit Hilfe von zwei Grundbeziehungen zu bestimmen.

Die erste Beziehung erhält man aufgrund der Tatsache, daß der Markrozustand, gebildet aus allen Zuständen der Kolonne c, sich im statistischen Gleichgewicht mit dem Zustand $P^a_{S_1-1}$ befinden muß (Bild 8.3).

Die erste Beziehung lautet deshalb:

$$P^a_{S_1-1} = \frac{\nu}{\lambda} \cdot \sum_{i=0}^{S} P_i^c \quad . \tag{8.8}$$

Die zweite Beziehung ist durch die Normierungsbedingung des Zustandsraumes gegeben

$$\sum_{i=0}^{S_1-1} P_i^a + \sum_{i=0}^{S} P_i^b + \sum_{i=0}^{S} P_i^c = V \quad . \tag{8.9}$$

Die endgültigen Beziehungen zur Bestimmung von P_X^1 und P_X^2 erhält man schließlich aus den Gl.(8.8) und (8.9) durch entsprechendes Einsetzen gemäß Gl.(8.7b):

$$P_X^2 = \underbrace{\frac{\nu}{\lambda} \cdot \sum_{i=0}^{S} A_i^c}_{C1} \cdot P_X^1 + \underbrace{\frac{\nu}{\lambda} \cdot \sum_{i=0}^{S} B_i^c}_{C2} \cdot P_X^2 \tag{8.10}$$

$$V = \underbrace{\left[\sum_{i=0}^{S_1-1} A_i^a + \sum_{i=0}^{S} (A_i^b + A_i^c) \right]}_{C3} \cdot P_X^1 + \underbrace{\left[\sum_{i=0}^{S_1-1} B_i^a + \sum_{i=0}^{S} (B_i^b + B_i^c) \right]}_{C4} \cdot P_X^2$$

Unter Einbeziehung der eingeführten Kurzbezeichnungen für die Summen erhält man nach einigen algebraischen Umformungen die Werte für die a-priori-Wahrscheinlichkeiten, aus denen sich alle Zustandswahrscheinlichkeiten gemäß Gl.(8.7a) bestimmen lassen:

$$P_X^1 = \frac{(1-C2) \cdot V}{(1-C2) \cdot C3 + C1 \cdot C4}$$

$$P_X^2 = \frac{C1 \cdot V}{(1-C2) \cdot C3 + C1 \cdot C4} \tag{8.11}$$

c) Charakteristische Verkehrsgrößen

Aus den Zustandswahrscheinlichkeiten - stationär sowie transient - erhält man die charakteristischen Größen zur Beurteilung der Systemparameter:

- die Wahrscheinlichkeit, daß Anforderungen am Schalter (SWitch) abgewiesen werden

$$B_{SW}(t) = \sum_{i=0}^{S} P_i^b(t) \quad , \tag{8.12}$$

- die Wahrscheinlichkeit, daß Anforderungen am System (SYStem) abgewiesen werden

$$B_{SYS}(t) = P_S^c(t) \quad , \tag{8.13}$$

- die mittlere Systembelastung

$$E[X(t)] = \sum_{i=0}^{S_1-1} i \cdot P_i^a(t) + \sum_{i=0}^{S} i \cdot P_i^b(t) + \sum_{i=0}^{S} i \cdot P_i^c(t) \quad , \tag{8.14}$$

- die Wahrscheinlichkeit, daß sich das System in der Schalterstellung AUS im Leerzustand befindet:

$$P_{LC}(t) = P_O^b(t) \quad , \tag{8.15}$$

dieses kann als Maß für verlorene Systemkapazität (Lost Capacity) gedeutet werden.

8.3.2 Die intermittierende Regelung in ihrer allgemeinen Form

a) Zustandsdiagramm

Im vorangehenden Abschnitt wurde vorausgesetzt, daß von jedem Meldungstyp höchstens ein einziger gleichzeitig unterwegs ist.

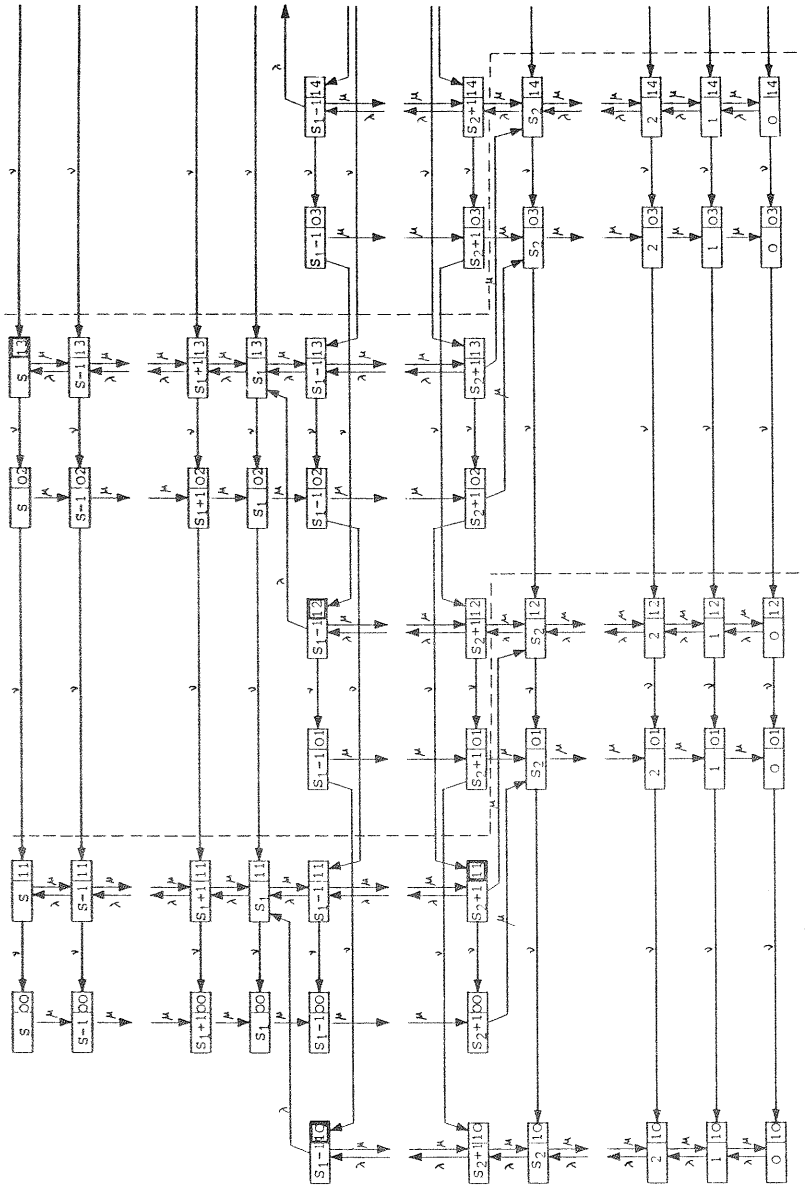


Bild 8.4: Zustandsdiagramm (Intermittierende Regelung, allgemeine Form).

Bedingt durch die statistischen Fluktuationen können aber bei einem zu eng gewählten Abstand zwischen beiden Schwellenwerten S_1 und S_2 mehrere Meldungen des gleichen Typs überlappend gesendet werden.

Für eine diesbezügliche Untersuchung wird das Zustandsdiagramm nach Bild 8.4 betrachtet. Dieses Diagramm ist komplizierter. Dennoch ist nach einer genaueren Betrachtung die Struktur klar erkennbar. Zuerst soll gezeigt werden, daß das im Abschnitt 8.3.1 behandelte einfachere Zustandsdiagramm in diesem Diagramm enthalten ist. Dazu werden alle Zustände rechts von der linken gestrichelten Linie vernachlässigt, so daß auch alle Übergänge aus diesen Zuständen wegfallen. Der einzige Übergang zum vernachlässigten Zustandsraum, nämlich $S_2\,|\,12 \rightarrow S_2{+}1\,|\,12$, wird in den Übergang $S_2\,|\,12 \rightarrow S_2{+}1\,|\,11$ umgewandelt. Wird anschliessend noch der verbleibende Teil der Zustandskolonnen (0,1) und (1,2) nach links verschoben, so erhält man genau Bild 8.3.

Die Berücksichtigung mehrerer gleichzeitiger Meldungen führt dazu, daß der Zustandsraum sich über den soeben betrachteten Zustandsübergang hinweg fortsetzt in einen entsprechenden zweiten Teil-Zustandsraum wiederum in einen dritten Teil-Zustandsraum usw. In jedem neu hinzukommenden Teil-Zustandsraum werden jeweils ein Meldungspaar (EIN/AUS) zusätzlich betrachtet. Nach Berücksichtigung aller Teil-Zustandsräume muß das Gesamt-Zustandsdiagramm ähnlich wie oben beschrieben, abgegrenzt werden. Ferner sind alle Übergänge, die das Eintreffen einer Meldung betreffen dadurch charakterisiert, daß die Zustandskennung sich entsprechend ändert: die Schalterstellung wird gewechselt und die Anzahl der Meldungen, die unterwegs sind, verringert sich um eins.

b) Zustandswahrscheinlichkeiten

Wie im einfacheren Modell können auch hier die Zustandswahrscheinlichkeiten nach einem rekursiven Verfahren berechnet werden. Für die Durchführung sind für jeden hinzukommenden Teilzustandsraum zwei zusätzliche Rekursions-Startpunkte erforderlich. Beschränkt man sich darauf, daß von jedem Meldungstyp höchstens zwei zugelassen sind, so werden vier a-priori-Wahrscheinlichkeiten benötigt:

$$P_X^1 = P_{S_1-1}^{10} \quad , \quad P_X^2 = P_{S_2+1}^{11} \quad , \quad P_X^3 = P_{S_1-1}^{12} \quad \text{und} \quad P_X^4 = P_S^{13} \tag{8.16}$$

Da die Zustände in derselben Zustandskolonne alle die gleiche Kennung haben, wird für dieses Modell keine spezielle Kolonnenschreibweise benutzt. Wie im Abschnitt 8.3.1 wird das System von rekursiven Gleichungen aus dem Zustandsdiagramm abgeleitet. Aus drei Beziehungen zwischen den Zustandswahrscheinlichkeiten und der Normierungsbedingung bestimmt man die a-priori-Wahrscheinlichkeiten, so daß anschließend alle Zustandswahrscheinlichkeiten aus deren Linearkombination bestimmt werden können. Für gewisse Parameterbereiche war dieser rekursive Algorithmus numerisch nicht stabil und für diese Fälle wurde die iterative Methode verwendet.

c) Charakteristische Verkehrsgrößen

Aus den Zustandswahrscheinlichkeiten, die hier nur stationär betrachtet werden, sind die charakteristischen Verkehrsgrößen wie folgt zu bestimmen:

- die Wahrscheinlichkeit, daß Anforderungen am Schalter abgewiesen werden. (Summe sämtlicher Zustandswahrscheinlichkeiten, die zu einem offenen Schalter gehören)

$$B_{SW} = \sum_{i=S_2+1}^{S} \left[P_i^{00} + P_i^{02} + \ldots \right] + \sum_{i=0}^{S_1-1} \left[P_i^{01} + P_i^{03} + \ldots \right] \tag{8.17}$$

- die Wahrscheinlichkeit, daß Anforderungen am System abgewiesen werden. (Summe sämtlicher Zustandswahrscheinlichkeiten, die zu einem geschlossenen Schalter und einem vollem System gehören)

$$B_{SYS} = P_S^{11} + P_S^{13} + \ldots . \tag{8.18}$$

- die mittlere Systembelastung

$$E[X] = \sum_{i=0}^{S_1-1} i \cdot \left[P_i^{10} + P_i^{12} + \ldots + P_i^{01} + P_i^{03} + \ldots \right] + \sum_{i=S_2+1}^{S} i \cdot \left[P_i^{00} + P_i^{02} + \ldots + P_i^{11} + P_i^{13} + \ldots \right] \tag{8.19}$$

- die Wahrscheinlichkeit, daß sich das System in der Schalterstellung AUS im Leerzustand befindet (Maß für verlorene Kapazität - Lost Capacity)

$$P_{LC} = P_O^{01} + P_O^{03} + \ldots . \tag{8.20}$$

- die Wahrscheinlichkeit, daß mehrere Meldungen gleichzeitig unterwegs sind (Summe sämtlicher Zustandwahrscheinlichkeiten, die zu Zuständen mit mehr als einer Meldung gehören)

$$P\{m>1\} = \sum_{i=0}^{S_1-1} \left[P_i^{12} + P_i^{14} + \ldots + P_i^{03} + P_i^{05} + \ldots \right] + \sum_{i=S_2+1}^{S} \left[P_i^{13} + P_i^{15} + \ldots + P_i^{02} + P_i^{04} + \ldots \right] . \tag{8.21}$$

8.3.3 Die Wechselregelung

a) Zustandsdiagramm

Als nächstes wird das im Bild 8.5 dargestellte Zustandsdiagramm für das Verkehrsmodell mit einer Wechselregelung (Bild 8.2b) betrachtet. Geregelt wird zwischen der normalen Ankunftsrate λ_1 und der gedrosselten Ankunftsrate $\lambda_2 < \lambda_1$.

Analog zum Abschnitt 8.3.1 werden die Zustände durch die folgenden Kennungen charakterisiert:

(1,k) : normaler Betrieb (Ankunftsrate λ_1)
(2,k) : gedrosselter Betrieb (Ankunftsrate λ_2)
k = 0,1,2 Meldungen unterwegs.

Da die Zustandskennungen in einer Zustandskolonne teilweise unterschiedlich sind, wird die in Bild 8.5 definierte Kolonnenschreibweise benutzt.

Das Zustandsdiagramm zeichnet sich dadurch aus,

- daß die Struktur in hohem Maße symmetrisch ist,
- daß die beiden inneren Zustandskolonnen b und c über jeweils nur einen einzigen Übergang verlassen werden können, und zwar in Kolonne b über den oberen Schwellenwert S_1-1 und in Kolonne c über den unteren Schwellenwert S_2+1.

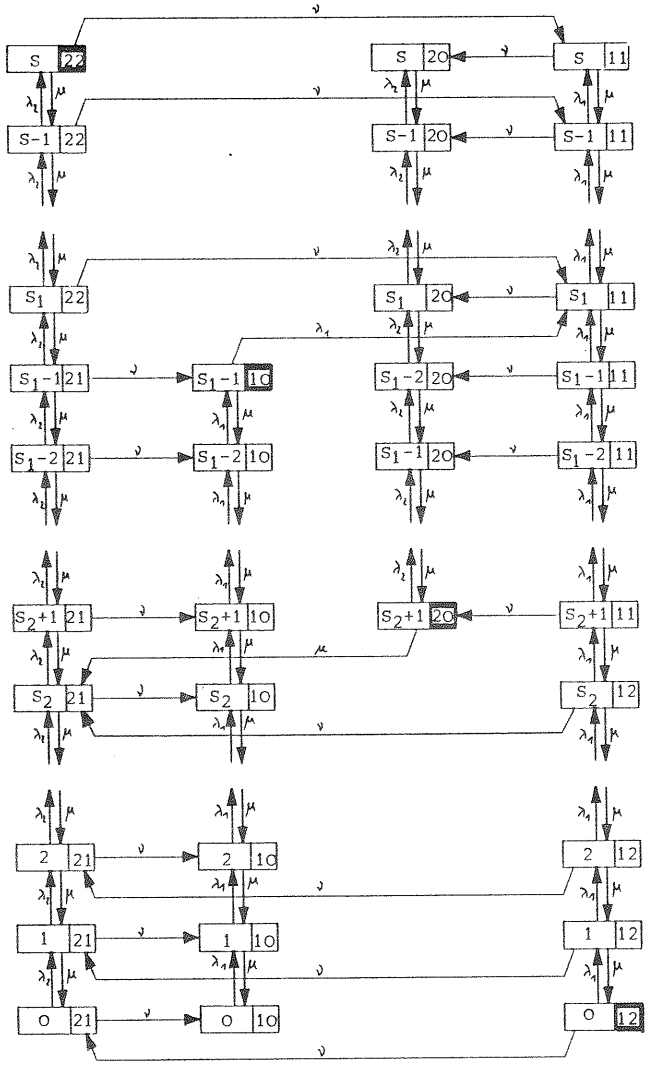


Bild 8.5: Zustandsdiagramm (Wechselregelung).

- daß in den beiden äußeren Zustandskolonnen a und d von jedem Zustand aus ein Kolonnenwechsel stattfinden kann, entweder in die benachbarte innere Kolonne oder in die gegenüberliegende äußere Kolonne.

Dieses Zustandsdiagramm geht als Spezialfall mit $\lambda_2 = 0$ in das Zustandsdiagramm für die intermittierende Regelung über (Bild 8.3). In diesem Fall sind nämlich alle Zustände in Kolonne a oberhalb des Zustandes $(S_2,2,1)$ nicht vorhanden und fallen deshalb zusammen mit den Übergängen weg. Verschiebt man jetzt den verbleibenden unteren Teil dieser Kolonne a unter Kolonne c und berücksichtigt man gleichzeitig, daß alle Übergänge für λ_2 nicht vorhanden sind, erhält man Bild 8.3.

b) Zustandswahrscheinlichkeiten

Die Zustandswahrscheinlichkeiten sind auch in diesem Falle rekursiv lösbar. Dazu ist in jeder Kolonne eine a-priori-Wahrscheinlichkeit erforderlich (Bild 8.5):

$$P_X^1 = P_S^a \;,\; P_X^2 = P_{S_1-1}^b \;,\; P_X^3 = P_{S_2+1}^c \quad \text{und} \quad P_X^4 = P_O^d \tag{8.22}$$

Die rekursiven Gleichungen und die zusätzlich benötigten Beziehungen lassen sich in ähnlicher Weise wie in Abschnitt 8.3.1 ableiten.

c) Charakteristische Verkehrsgrößen

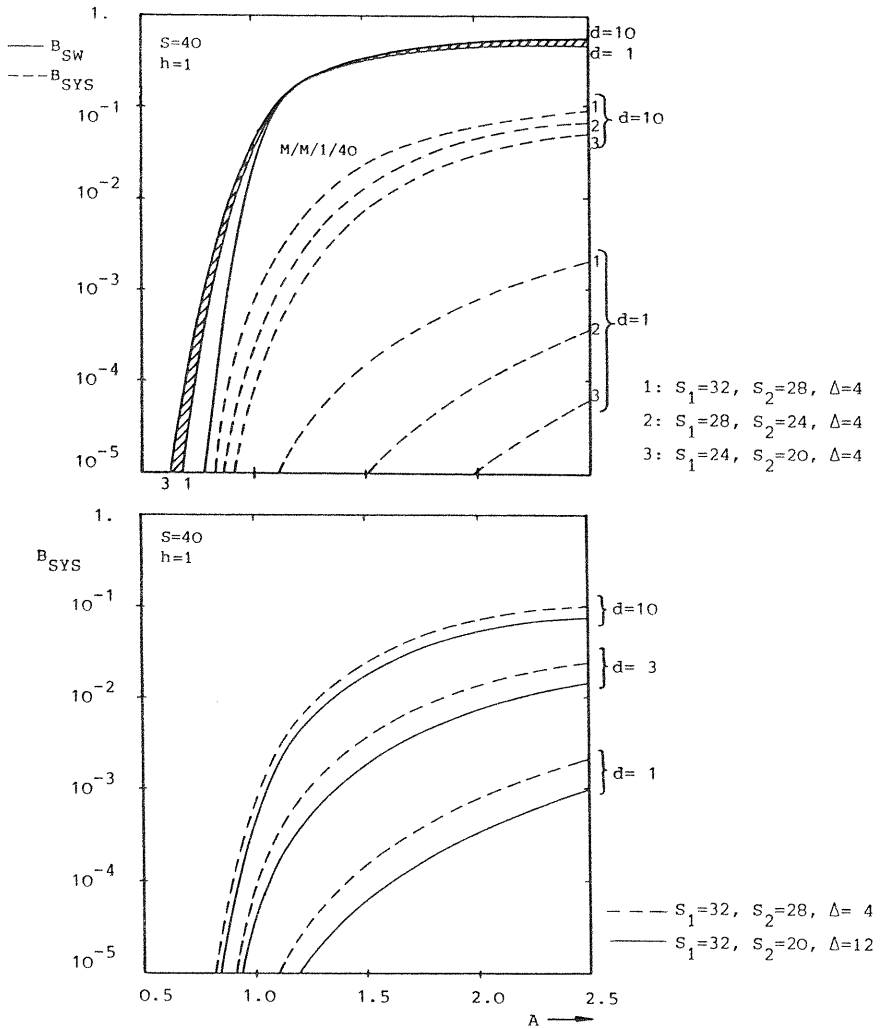
Als charakteristische Verkehrsgrößen werden für die Wechselregelung betrachtet:

- die Verlustwahrscheinlichkeit am System (Zeitblockierung)

$$B_{SYS}(t) = P_S^a(t) + P_S^c(t) + P_S^d(t) \tag{8.23}$$

- die Verlustwahrscheinlichkeit für eine ankommende Anforderung, die durch entsprechende Gewichtung der Zustandswahrscheinlichkeiten mit den Ankunftsraten ermittelt wird.

$$B_{ANF}(t) = \frac{\lambda_1(t) \cdot P_S^d(t) + \lambda_2(t) \cdot [P_S^a(t) + P_S^c(t)]}{\lambda_1(t) \cdot \left[\sum_{i=0}^{S_1-1} P_i^b(t) + \sum_{i=0}^{S} P_i^d(t)\right] + \lambda_2(t) \cdot \left[\sum_{i=0}^{S} P_i^a(t) + \sum_{i=S_2+1}^{S} P_i^c(t)\right]} \tag{8.24}$$



Intermittierende Regelung: Verlustwahrscheinlichkeit am Schalter B_{SW} bzw. am System B_{SYS} als Funktion vom Verkehrsangebot A.

Bild 8.6: a) Vergleich mit Warteschlangensystem ohne Zweipunktregelung.
b) Verschiedene mittlere Verzögerungen (d=1,10).
c) Verschiebung der beiden Schwellenwerte S_1 und S_2 bei konstantem Abstand Δ = 4.

Bild 8.7: a) Verschiedene mittlere Verzögerungen (d=1,10).
b) Unterschiedlicher Abstand zwischen beiden Schwellenwerten S_1 und S_2 bei festem Wert S_1 = 32; (Δ = 4,12).

- die mittlere Systembelastung

$$E[X(t)] = \sum_{i=0}^{S} i \cdot P_i^a(t) + \sum_{i=0}^{S_1-1} i \cdot P_i^b(t) + \sum_{i=S_2+1}^{S} i \cdot P_i^c(t) + \sum_{i=0}^{S} i \cdot P_i^d(t) \qquad (8.25)$$

- der Durchsatz D(t)

$$D(t) = \mu \cdot [1 - P_0^a(t) - P_0^b(t) - P_0^d(t)] \quad . \qquad (8.26)$$

8.4 Numerische Ergebnisse

Anhand von numerischen Beispielen werden in diesem Abschnitt einige interessante Ergebnisse zusammengestellt.

8.4.1 Stationäre Ergebnisse

Betrachtet wird ein Warteschlangensystem mit Kapazität S = 40, das mit Hilfe der intermittierenden Regelung gegen Überlast geschützt werden soll. Die mittlere Bedienungszeit sei h = 1/μ = 1.

a) Verlustwahrscheinlichkeit

Bild 8.6 zeigt für verschiedene Parameter die Verlustwahrscheinlichkeit B_{SW} bzw. B_{SYS} in Abhängigkeit des angebotenen Verkehrs A = λ/μ. Als Referenz ist auch die Kurve für die Verlustwahrscheinlichkeit des gleichen Systems ohne Regelung angegeben. Die Parameter werden in zweierlei Hinsicht variiert:

- einerseits werden verschiedene Werte für den oberen Schwellenwert S_1 = 32, 28, 24 gewählt und damit gleichzeitig der Wert für den unteren Schwellenwert S_2 festgelegt (Differenz Δ = 4),
- andererseits werden zwei Werte für die Verzögerung betrachtet: d = 1,10.

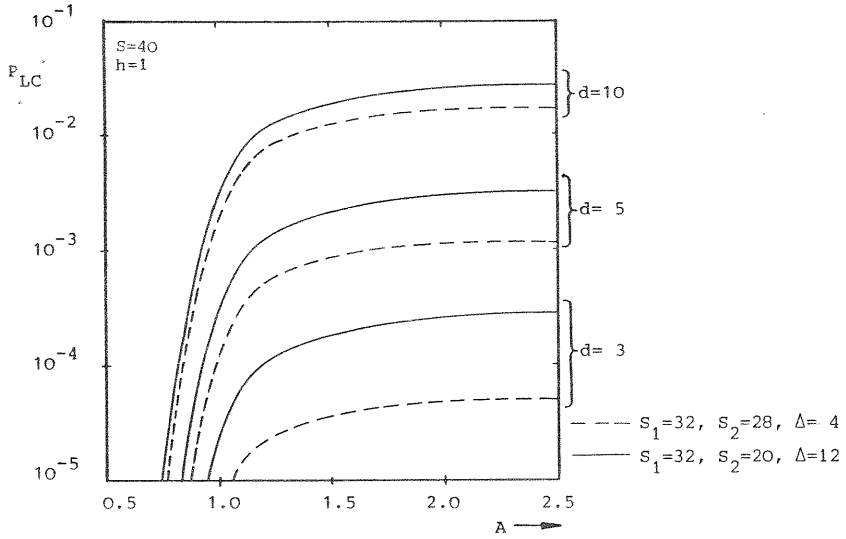
Aus dem Verlauf der Kurven können einige Feststellungen gemacht werden:

- die Charakteristik der Überlastabwehr am Schalter, gegeben durch B_{SW}, deckt sich mit der Verlustkurve des Referenzsystems im Überlastbereich,
- die Kurven für B_{SW} sind ziemlich unempfindlich gegen Parameteränderungen: bei niedrigen Verkehrswerten ist die Schwellen-

wertverschiebung, bei hohen Verkehrswerten die Verzögerung für den Kurvenverlauf verantwortlich,

- durch die Schutzwirkung müssen teilweise beträchtlich weniger Anforderungen am System selbst (Kurven B_{SYS}) abgewiesen werden,
- erwartungsgemäß ist das Warteschlangensystem umso besser geschützt, je kleiner die Verzögerung ist,
- die Verlustwahrscheinlichkeit B_{SYS} reagiert empfindlicher auf eine Verschiebung des oberen Schwellenwertes S_1, wenn die Verzögerung klein ist (d = 1).

In Bild 8.7 ist die Verlustwahrscheinlichkeit B_{SYS} nochmals dargestellt. Dieses Diagramm zeigt, daß B_{SYS} nur geringfügig verbessert werden kann, wenn der Abstand zwischen den beiden Schwellenwerten, bei einem fixierten oberen Schwellenwert $S_1 = 32$, von $\Delta = 4$ auf $\Delta = 12$ vergrößert wird.



Intermittierende Regelung: Wahrscheinlichkeitsmaß für verlorene Systemkapazität P_{LC} als Funktion vom Verkehrsangebot A.

Bild 8.8: a) Verschiedene mittlere Verzögerungen (d=1,10).
b) Unterschiedlicher Abstand zwischen beiden Schwellenwerten S_1 und S_2 bei festem Wert $S_1 = 32$; ($\Delta = 4,12$).

b) Wahrscheinlichkeitsmaß für ungenutzte Kapazität

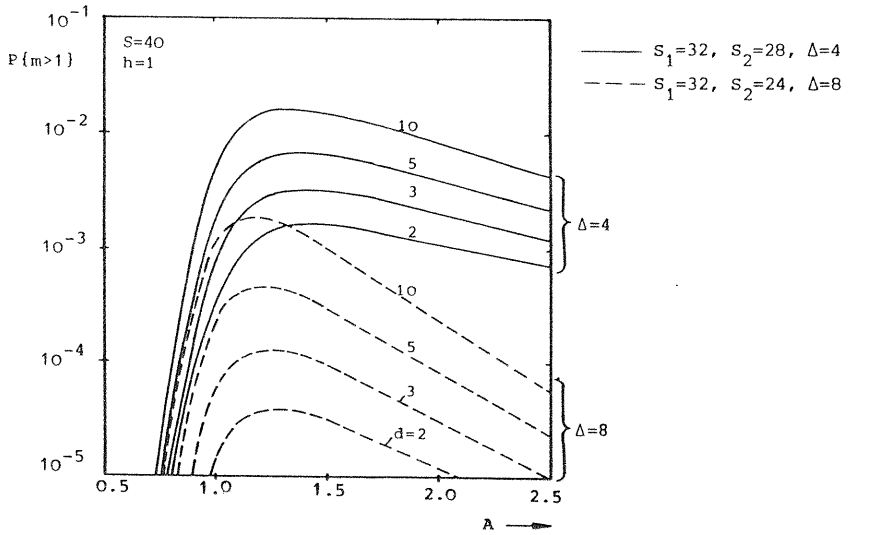
Wie aus Bild 8.8 hervorgeht, führt eine Verringerung des unteren Schwellenwertes S_2 bei festem Wert von S_1 dazu, daß mit einer größeren Wahrscheinlichkeit das System sich vollständig entleeren kann, bevor der Schalter wieder geschlossen (Stellung EIN) wird. Es geht deshalb mit einer größeren Wahrscheinlichkeit Kapazität verloren, was sich durch höhere Werte von P_{LC} ausdrückt.

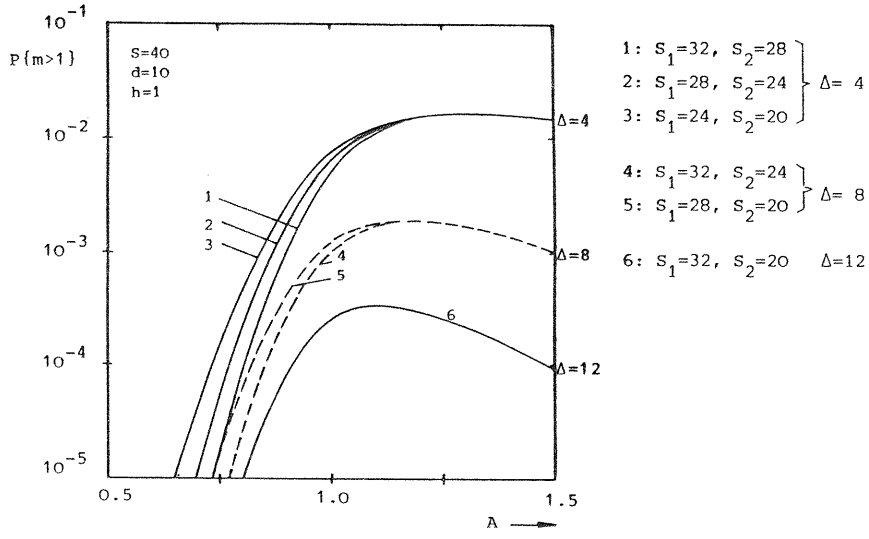
c) Wahrscheinlichkeit für mehrere gleichzeitige Meldungen

Als nächstes wird die Wahrscheinlichkeit, daß mehrere Meldungen gleichzeitig unterwegs sind, P{m>1} betrachtet. An dieser Stelle sei erwähnt, daß die Berechnung der vorherigen Resultate sowohl mit dem einfachen (Bild 8.3) als auch mit dem komplexen Zustandsdiagramm (Bild 8.4) durchgeführt worden ist. In dem untersuchten Parameterbereich wurde eine sehr gute Übereinstimmung festgestellt.

In Bild 8.9 werden zwei Werte für den Abstand zwischen den beiden Schwellenwertgrößen S_1 und S_2 betrachtet: $\Delta = 4$ und $\Delta = 8$. Außerdem werden verschiedene Werte d für die Verzögerung gewählt. Aus den beiden Kurvenscharen geht hervor, daß je größer der Abstand zwischen den Schwellenwertgrößen und je kleiner die Verzögerung ist , die Wahrscheinlichkeit für mehrere überlappende Meldungen desto geringer ist. Ferner weisen die Kurven ein Maximum im Verkehrsbereich A = 1.0 bis 1.5 auf, denn in diesem Bereich findet der Regelungsvorgang besonders häufig statt. Im niedrigeren Verkehrsbereich ist der Schalter vorwiegend in der Stellung EIN, für hohe Verkehrswerte ist er größtenteils in der Stellung AUS.

In Bild 8.10 ist der interessante Bereich um den Verkehrswert A = 1.0 nochmals hervorgehoben. Betrachtet werden die Schwellenwertabstände $\Delta = 4$, 8 und 12 für eine mittlere Verzögerung d = 10. Zudem werden die beiden Schwellenwerte - bei gleichbleibendem Abstand - noch verschoben. Wie aus den Kurven entnommen werden kann, wird P{m>1} im Hochlastbetrieb nur vom Abstand zwischen beiden Schwellenwerten bestimmt. Die Lage dieses Schwellenwertpaares bei konstantem Abstand spielt jedoch bei Verkehrswerten





Intermittierende Regelung: Wahrscheinlichkeit für mehrere gleichzeitige Meldungen P{m>1} als Funktion vom Verkehrsangebot A.

Bild 8.9: a) Verschiedene mittlere Verzögerungen (d=2,3,4,10).
b) Unterschiedlicher Abstand zwischen beiden Schwellenwerten S_1 und S_2 bei festem Wert S_1=32; (Δ=4,8).

Bild 8.10: a) Verschiebung der beiden Schwellenwerte S_1 und S_2 bei konstantem Abstand Δ=4,8.
b) Unterschiedlicher Abstand zwischen beiden Schwellenwerten S_1 und S_2 bei festem Wert S_1=32; (Δ=4,8,12).

A<1.0 eine Rolle. Ihr Einfluß ist bei kleinem Schwellenwertabstand am größten.

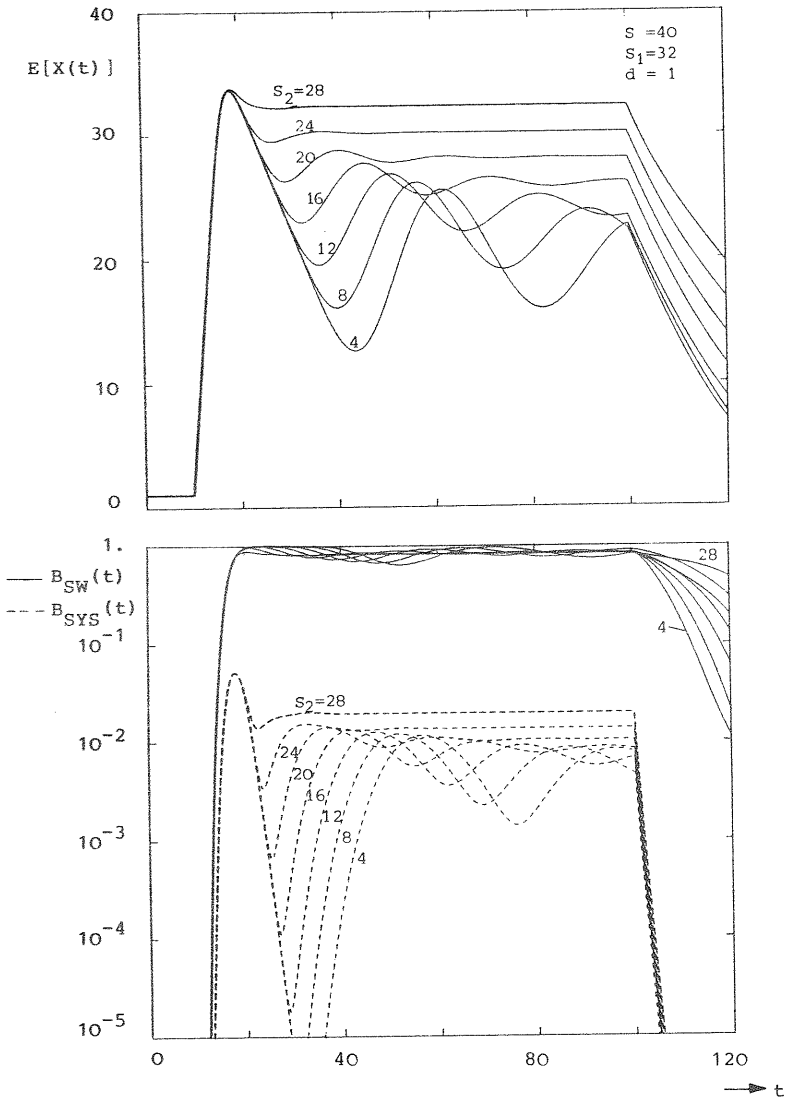
8.4.2 Transiente Ergebnisse

Im folgenden wird das dynamische Systemverhalten analysiert, wobei zuerst die intermittierende Regelung und danach die Wechselregelung betrachtet wird. Untersucht wird wieder ein Warteschlangensystem mit Kapazität S = 40. Die mittlere Bedienungszeit h = 1/μ = 1 sei gleichzeitig die Zeiteinheit. Als Überlastimpuls sei eine Ankunftsrate λ(t) betrachtet, die sich gemäß einer rechteckförmigen Funktion ändert: stationäre Ankunftsrate λ = 0.5, Überlastintensität λ_{max} = 6, Überlastbeginn t = 10, Überlastende t = 100.

8.4.2.1 Intermittierende Regelung

a) mittlere Systembelastung

Bild 8.11 zeigt den Verlauf der mittleren Systembelastung E[X(t)], wenn bei einem oberen Schwellenwert S_1 = 32 und bei einer mittleren Verzögerung d = 1, der untere Schwellenwert S_2 variiert wird. Durch die schlagartige Vergrößerung der Ankunftsrate λ(t) zum Zeitpunkt t = 10 steigt die Systembelastung schnell bis auf den oberen Schwellenwert S_1 an (Anstiegsrate 5 Anforderungen pro Zeiteinheit). Wird S_1 erreicht (t = 16.4), so wird eine Meldung zum Schalter gesendet,um den Ankunftsprozeß abzuschalten. Bis zur Abschaltung, also während der Verzögerung d = 1, kann die Systembelastung aber noch weiter ansteigen. Das Maximum liegt bei 34.5 und wird aus nachher angegebener Begründung etwas später erreicht als erwartet: t = 18.3. Sobald aber der Schalter in der Stellung AUS steht, sinkt die Belastung mit der Bedienungsrate μ = 1 wieder ab. Beim Erreichen des unteren Schwellenwertes S_2 wird wiederum eine Meldung zum Einschalten gesendet. Betrachtet man zum Beispiel S_2 = 4, so wird festgestellt, daß die mittlere Systembelastung nur bis E[X(t)] = 12.9 absinkt (t = 44). Die Tangente für die Abnahme der Belastung würde E[X(t)] = 4 zum Zeitpunkt t = 18.3 + (34.5-4) = 48.8 erreichen. Dies hängt mit dem zeitlichen Verlauf der stochastischen Prozesse für Ankunft, Bedienung und Verzögerung zusammen. Durch



Intermittierende Regelung: Kurze mittlere Verzögerungszeit d = 1, Überlastdauer T = 90.

Bild 8.11: Mittlere Systembelastung E[X(t)].

Bild 8.12: Verlustwahrscheinlichkeit am Schalter $B_{SW}(t)$ bzw. am System $B_{SYS}(t)$. Verkehrsparameter: $\lambda(\infty) = 0.5$, $\lambda_{MAX} = 6.0$, $h = 1$.

ihre zeitliche Abhängigkeit kann der Verlauf von E[X(t)] nicht mehr mit einfachen Mittelwertüberlegungen vorausgesagt werden. Auch dieses Beispiel zeigt die Bedeutung der transienten Analyse.

Die Kurven zeigen, daß das Ein- und Ausschalten zu gedämpften Schwingungen führen kann. Dabei ist die Schwingungsamplitude umso größer, je kleiner der untere Schwellenwert S_2 gewählt wird. Bei größeren Anfangsamplituden wird sogar am Ende eines langen Überlastimpulses noch kein eingeschwungener Zustand erreicht. Es zeigt sich außerdem, daß die Wahl eines niedrigen Schwellenwertes S_2 auch einen niedrigen Endwert für die Systembelastung ergibt. Entsprechend hoch ist dann aber auch die Wahrscheinlichkeit für Kapazitätsverlust P_{LC}, die ein ähnliches Schwingungsverhalten aufweist.

b) Verlustwahrscheinlichkeit

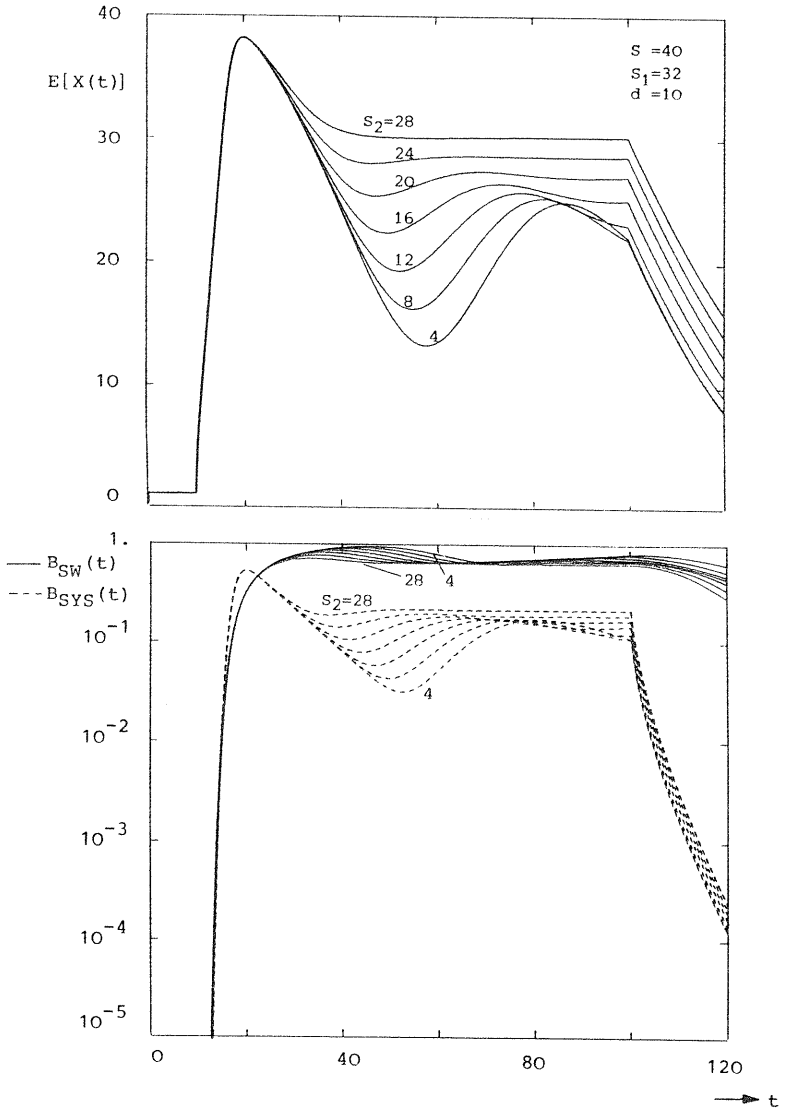
Bild 8.12 gibt Aufschluß über den entsprechenden Verlauf der Verlustwahrscheinlichkeiten am Schalter $B_{SW}(t)$ bzw. am System $B_{SYS}(t)$. Der Systembelastung entsprechend erreicht auch $B_{SYS}(t)$ bei erster Abschaltung den Höchstwert. Danach folgt, je nach Wert von S_2 ein ausgeprägtes gedämpftes Schwingungsverhalten. Der Verlauf von $B_{SW}(t)$ ist ebenfalls schwingungsbehaftet, jedoch mit wesentlich geringerer Amplitude. Es sei auch auf die für überlastete Systeme typische, lange Erholungszeit hingewiesen. Durch die sich langsam abbauende Systembelastung weist $B_{SW}(t)$ auch nach Ende der Überlast noch einige Zeit hohe Werte auf (der Wert 10^{-5} wird bei etwa t = 180 erreicht).

c) Systemverhalten bei großer Verzögerung

Die Bilder 8.13 und 8.14 geben die entsprechenden Kurven für die wesentlich längere mittlere Verzögerung d = 10.

Deutlich erkennt man die schlechtere Regelwirkung:

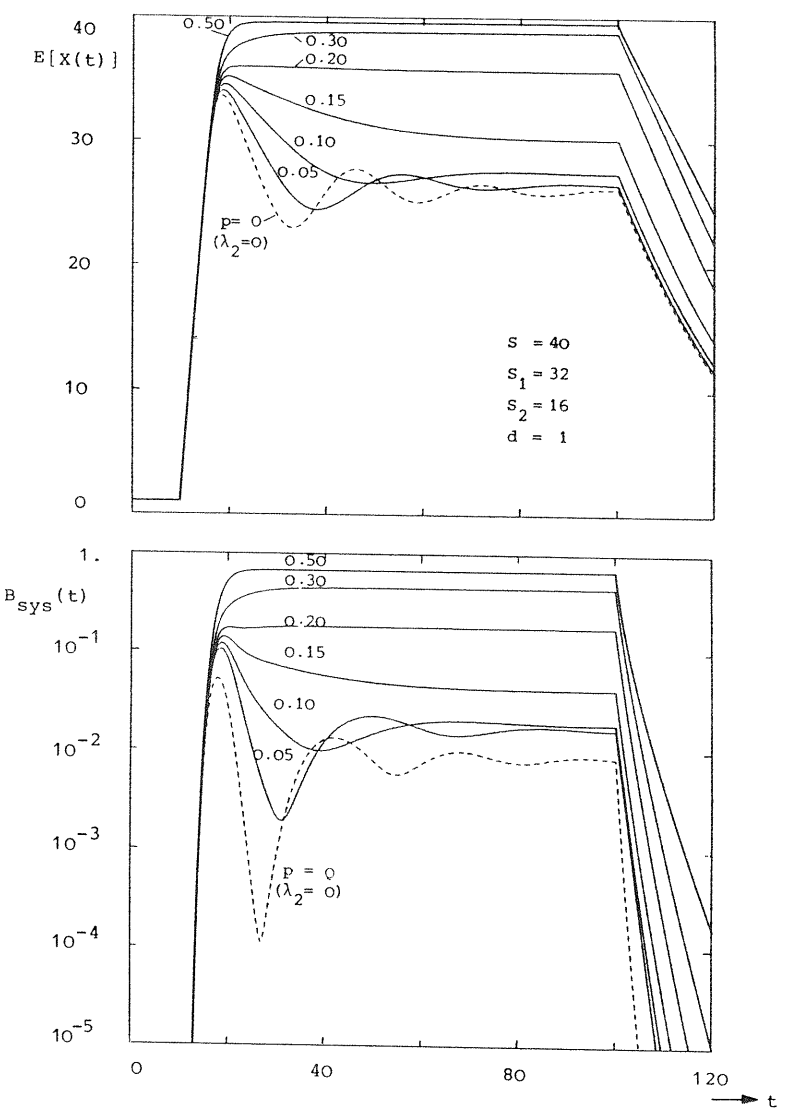
- die mittlere Systembelastung E[X(t)] erreicht bis zum Ausschalten einen höheren Wert,
- nach dem Einschwingen stellt sich jeweils eine geringere Systembelastung ein und somit ist $P_{LC}(t)$ auch größer,



Intermittierende Regelung: lange mittlere Verzögerungszeit d = 10, Überlastdauer T = 90.

Bild 8.13: Mittlere Systembelastung E[X(t)].

Bild 8.14: Verlustwahrscheinlichkeit am Schalter $B_{SW}(t)$ bzw. am System $B_{SYS}(t)$. Verkehrsparameter: $\lambda(\infty) = 0.5$, $\lambda_{MAX} = 6.0$, $h = 1$.



Wechselregelung: Einfluß verschiedener Drosselungsfaktoren p bei kurzer mittlerer Verzögerungszeit d = 1, Überlastdauer T = 90.

Bild 8.15: Mittlere Systembelastung E[X(t)].

Bild 8.16: Verlustwahrscheinlichkeit am System $B_{SYS}(t)$; (Zeitblockierung). Verkehrsparameter: $\lambda(\infty) = 0.5$, $\lambda_{MAX} = 6.0$, $h = 1$.

- ein wesentlich höherer Anteil von Anforderungen wird anstatt am Schalter, am System abgewiesen.

Ferner haben die Schwingungen eine größere Periode und eine kleinere Amplitude. Speziell bei $B_{SYS}(t)$ fehlt die extrem große Anfangsamplitude.

8.4.2.2 Wechselregelung

Das dynamische Verhalten bei einer entsprechenden Wechselregelung ist in den Bildern 8.15 und 8.16 veranschaulicht.

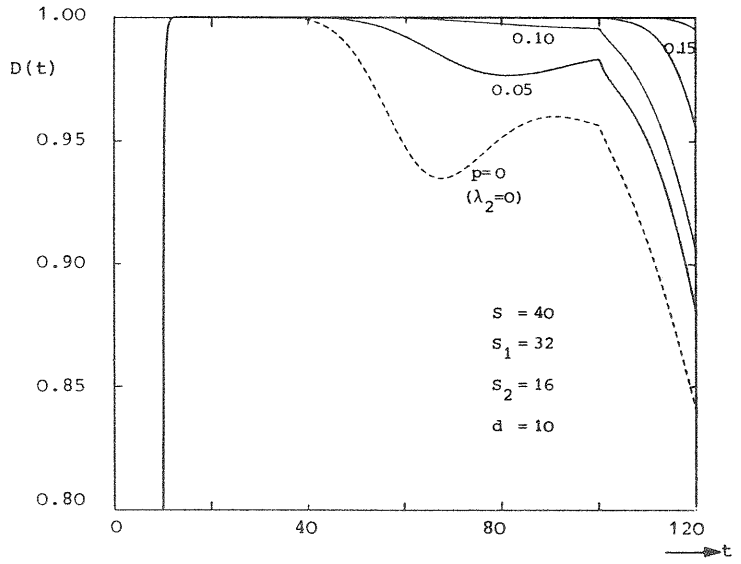
Wieder ist die Systemkapazität $S = 40$, für den oberen Grenzwert gilt $S_1 = 32$ und zusätzlich wird $S_2 = 16$ und $d = 1$ gewählt. Für die gedrosselte Verkehrsrate gilt $\lambda_2(t) = p \cdot \lambda_1(t)$.

Die gestrichelte Kurve stellt den Spezialfall der intermittierenden Regelung dar. Aus den Kurven geht hervor, daß bei Drosselung auf eine kleine Ankunftsrate, zum Beispiel 10 % des Normalwertes, die Schwingungen weitgehend vermieden werden können, ohne daß die Verlustwahrscheinlichkeit (Zeitblockierung) $B_{SYS}(t)$ und die mittlere Systembelastung $E[X(t)]$ dramatisch schlechter werden. Eine geringere Drosselung verursacht jedoch wesentlich höhere Werte für $B_{SYS}(t)$.

Bild 8.17 zeigt die Schwankungen im Durchsatz D(t) für verschiedene Drosselungsfaktoren p und eine mittlere Verzögerung $d = 10$. Auch diese Kurven lassen erkennen, daß sich das völlige Ausschalten des Verkehrsstromes leistungsmindernd auswirkt.

8.5 Schlußfolgerung

Die Ergebnisse zeigen, daß das Ziel, einen Netzknoten durch Abschalten oder Drosselung der Paketrate an der Netzübergabestelle gegen eine Überlast zu schützen, um so besser erreicht wird, je kleiner die mittlere Verzögerungszeit ist bis die Regelung einsetzt. Diese als Zweipunktregelung gestaltete Überlastabwehrstrategie dient jedoch immer als Ergänzung zu den viel schnelleren lokalen Maßnahmen.



Wechselregelung: Schwankungen im Durchsatz bei verschiedenen Drosselungsfaktoren p und längerer mittlerer Verzögerungszeit d = 10, Impulsdauer T = 90.

Bild 8.17: *Durchsatz D(t).*
Verkehrsparameter: $\lambda(\infty) = 0.5$, $\lambda_{MAX} = 6.0$, $h = 1$.

Ferner zeigt sich, daß bei Vorgabe der Systemkapazität S und der mittleren Verzögerungszeit d:

- der obere Schwellenwert S_1 möglichst klein gewählt werden sollte (wegen B_{SYS}),
- der untere Schwellenwert S_2 möglichst groß gewählt werden sollte (wegen P_{LC}),
- der Abstand zwischen beiden Schwellenwerten möglichst groß gewählt werden sollte (wegen P{m>1}).

Aufgrund dieser widersprechenden Forderungen ist eine optimale Parameterwahl nur unter Verwendung einer geeigneten Zielfunktion möglich.

9. ÜBERLASTABWEHR DURCH ADAPTIVE ZEITÜBERWACHUNG

Die Kommunikation zwischen Anwenderprozessen beruht auf Protokollen, die einen einwandfreien Informationsaustausch gewährleisten sollen. Um dies verwirklichen zu können, werden automatische Fehlerkorrekturverfahren eingesetzt, wobei ein Paketverlust oder ein fehlerhaftes Paket durch Paketwiederholung korrigiert wird. In Überlastfällen können Paketwiederholungen jedoch zu einer bedeutenden Verschlechterung der Netzsituation führen. Bei Überschreitung einer Netzgrenzlast - wie dies in Abschnitt 5.3 gezeigt wurde - kann dies sogar in eine lawinenartige Überflutung des Netzes mit Paketkopien entarten. In diesem Kapitel werden Strategien untersucht, um einen Zusammenbruch des Netzes durch diese zusätzliche Blindlast zu verhindern.

9.1 Allgemeines

Wie in Bild 9.1 dargestellt ist, findet die Fehlerkorrektur und die Zeitüberwachung typischerweise auf drei Ebenen statt:

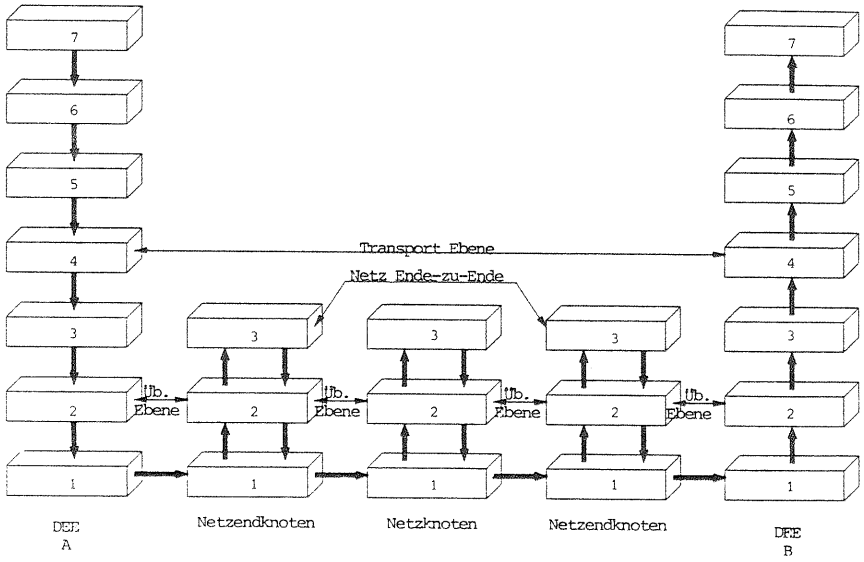


Bild 9.1: Hierarchische Implementierung der Fehlerkorrektur und der Zeitüberwachung (Time-Out Mechanismus).

- Übertragungsfehler und Abweisungen wegen eines Speicherüberlaufs werden jeweils abschnittsweise auf der Sicherungsebene behoben,
- fehlende Pakete werden entweder auf der Vermittlungsebene in den Netzendknoten oder auf der Transportebene bei den Netzbenutzern festgestellt und zur Wiederholung angefordert.

Obwohl die Fehlerkorrekturfunktion je nach Ebene andere Aufgaben zu erfüllen hat und deshalb verschieden implementiert wird, ist immer eine Zeitüberwachung vorhanden. Bei diesem sogenannten Time-Out-Mechanismus werden Pakete,deren Quittungen die Vorgabezeit überschritten haben, automatisch wiederholt. Der eingestellte Zeitbegrenzungswert auf der Vermittlungs- oder Transportebene kann die Verkehrseigenschaften des Netzes entscheidend beeinflussen:

- Ein zu langes Zeitintervall wird die Durchführung der Fehlerkorrektur verzögern, so daß beträchtliche Paketübermittlungszeiten und Speicherbelegungszeiten für Paketkopien entstehen können.
- Ein zu kurzes Zeitintervall führt zu unnötigen Paketwiederholungen, die eine zusätzliche Netzbelastung verursachen. Durch den Mitkopplungseffekt kann das Netz dann zusammenbrechen. Aus diesem Grund ist bei der Spezifikation von Überlastabwehrstrategien die Einstellung dieser Zeitüberwachung ebenfalls einzubeziehen.

In der Literatur wurde der Time-Out-Mechanismus schon in Verkehrsmodellen berücksichtigt. In [Lam (1976), Schweitzer/Lam (1976), Bux (1976)] wird bei der Untersuchung von Netzknotenmodellen die Paketübertragung aufgrund eines Wahrscheinlichkeitsparameters wiederholt. Der Einfluß des Time-Out-Intervalls auf ein "Send-and-Wait" Protokoll wird in [Fayolle/Gelenbe/Pujolle (1978)] analysiert. In [Sunshine (1977)] stellt sich heraus, daß die Einstellung der Zeitüberwachung zu den Parametern gehört, die die Wirksamkeit von Kommunikationsprotokollen entscheidend beeinflussen können. In [Kleinrock/Kermani (1980), Kermani/Kleinrock (1980)] gehört das Time-Out-Intervall zu den Grundparametern in einer analytischen Untersuchung über

die Datenflußsteuerung nach dem Fenstermechanismus. [Morris (1979)] behandelt dieses Thema unter allgemeinen Gesichtspunkten und identifiziert die Notwendigkeit für eine adaptive Einstellung der Zeitüberwachung. Das Phänomen der Instabilität bei Überschreitung einer Netzgrenzlast wird in [Butto/Colombo/Taggiasco/Tonietti(1977)] zum ersten Mal festgestellt.

9.2 Gesichtspunkte zur Durchführung der Zeitüberwachung

Zur Spezifikation einer geeigneten Zeitüberwachung sind folgende Aspekte zu berücksichtigen:

- Länge des Time-Out-Intervalls,
- Anzahl zugelassener Kopien pro Paket,
- Vorgehen bei einer Zeitüberschreitung.

9.2.1 Länge des Time-Out-Intervalls

Unterschieden wird zwischen einer festen und einer adaptiven Einstellung der Zeitüberwachung. Wie in Abschnitt 5.3 gezeigt wurde, ist eine Betriebsweise mit einem festen Time-Out-Intervall sehr empfindlich auf Fluktuationen in der Netzbelastung. Bereits eine vorübergehende hohe Quittierungsverzögerung kann eine Lawine von Paketkopien auslösen und ohne zusätzliche Maßnahmen würde das Netz sofort zusammenbrechen. Wird jedoch ein neugesetztes Time-Out-Intervall dem momentanen Netzzustand angepaßt, so wirkt sich der Time-Out-Mechanismus wesentlich geringer auf die Verkehrseigenschaften des Netzes aus.

Diesbezügliche Realisierungen unterscheiden sich in:
(Quantitative Aussagen sind in [van As (1983)] zu finden.)

- den verwendeten Informationen (Messung der Quittierungsverzögerung, Anzahl der Zeitüberschreitungen in einem Meßintervall, Überlastindikatoren),
- der Methode zur Bestimmung des Time-Out-Intervalls (Anzahl der Meßwerte, Länge des Meßintervalls, Gewichtung der Meßwerte, Berechnungsverfahren),
- der Update-Frequenz für Zustandsinformationen und Bestimmungsfrequenz für das Time-Out-Intervall (kontinuierlich, periodisch, lastabhängig).

9.2.2 Anzahl zugelassener Kopien pro Paket

Eine Begrenzung der Anzahl Zeitüberschreitungen pro Paket ist angemessen. Denn falls die Quittierungszeit eines Paketes mehrmals überschritten wird, kann davon ausgegangen werden, daß entweder irgendwo eine ernsthafte Überlastsituation im Übermittlungsweg aufgetreten ist, oder daß durch einen Hardware-Defekt (Netzknoten, Leitung, Empfänger) das Paket nicht vermittelt werden kann. Das Aussenden weiterer Paketkopien soll deshalb eingestellt werden,und der Zustand der logischen Verbindung soll zuerst mit Hilfe von Netzsteuerungspaketen überprüft werden.

9.2.3 Vorgehen bei einer Zeitüberschreitung: das Time-Out-Fenster

Der Netzzustand kann sich zwischen dem Zeitpunkt,zu dem das Time-Out gesetzt wird und dem Zeitpunkt, bei dem diese Zeitbegrenzung überschritten wird, entscheidend verändern. Deshalb ist eine adaptive Einstellung des Time-Out-Intervalls allein keine Garantie für ein stabiles Netzverhalten. Dieses trifft insbesondere für Überlastsituationen zu.

Als Maßnahme wird deshalb, in Analogie zu dem Fenstermechanismus bei der Datenflußsteuerung, das Konzept des Time-Out-Fensters eingeführt. Zur Gewinnung einer Steuergröße wird die Anzahl der Originalpakete in der betreffenden Verkehrsbeziehung - z.B. zwischen zwei Netzendknoten - kontinuierlich überwacht. Beim Eintreffen eines Paketes wird ein Zähler erhöht und er wird wieder erniedrigt bei Ankunft einer entsprechenden Quittung. Eine mögliche Prozeßrealisierung ist in Bild 9.2 dargestellt. Ereignet sich eine Zeitüberschreitung während einer Zeitspanne, in der sich der Zählerstand unterhalb eines Maximalwertes befindet, kann eine Paketkopie sofort generiert und abgeschickt werden. Der betreffende Zähler wird jedoch nicht erhöht. In diesem Fall ist das Time-Out-Fenster offen. Andererseits, wenn eine Zeitüberwachung anspricht während das Fenster geschlossen ist (schraffierter Bereich), so wird die Behandlung des Time-Outs verzögert,bis sich das Fenster wieder öffnet. Durch die Verzögerung wird einerseits das Paketvermittlungsnetz, das dann vermutlich bereits überlastet ist, keiner zusätzliche Belastung ausgesetzt und andererseits wird ein Teil der verzögerten

Time-Outs noch bevor ihre Behandlung stattfindet durch Eintreffen der entsprechenden Quittung zurückgesetzt.

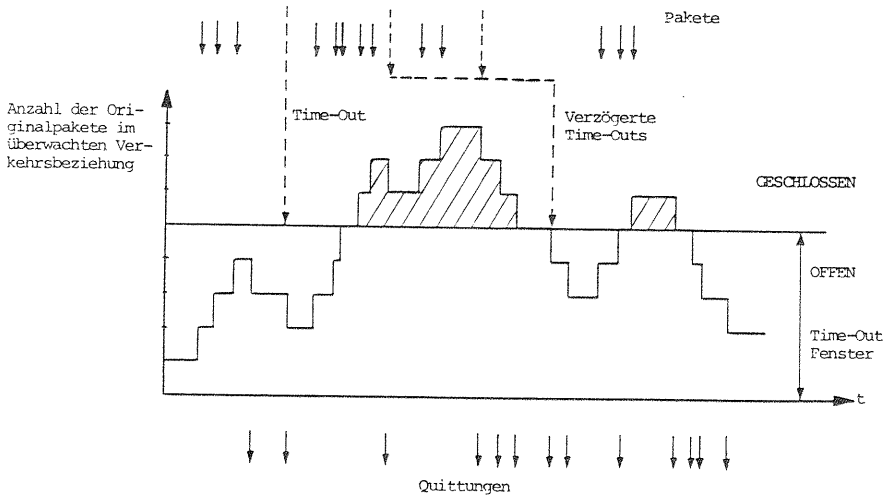


Bild 9.2: Das Time-Out Fenster.

9.2.4 Gesamtkonzept für den Time-Out-Mechanismus

In Bild 9.3 ist das Ablaufgeschehen für den Time-Out-Mechanismus schematisch dargestellt. Für jedes neue Paket wird eine Kopie abgespeichert und ein Time-Out gesetzt. Kopie und Time-Out werden wieder gelöscht, sobald die entsprechende Quittung eintrifft. Wird jedoch die eingestellte Zeitbegrenzung überschritten, bevor die Quittung zurückkommt, so wird beim offenen Time-Out-Fenster, das im Bild einem geschlossenen Schalter entspricht, sofort eine Paketkopie generiert und abgesendet. Gleichzeitig wird die neue Zeitbegrenzung nach den implementierten Regeln gesetzt. Anderenfalls wird bei geschlossenem Time-Out-Fenster (Schalter offen) die Behandlung des Time-Out-Ereignisses verzögert. Während dieser Verzögerungszeit kann beim Eintreffen der entsprechenden Quittung ein Time-Out und die zugehörige Paketkopie noch gelöscht werden.

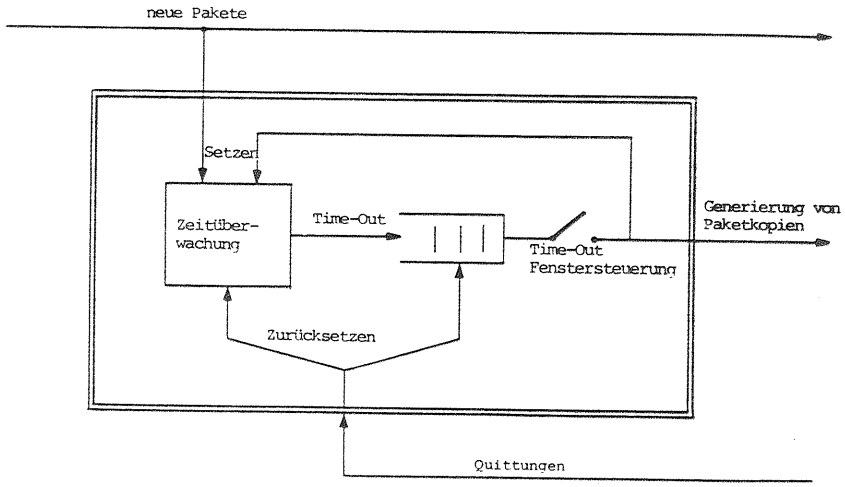


Bild 9.3: Gesamtkonzept für die Realisierung der Zeitüberwachung (Time-Out Mechanismus).

9.3 Modellbeschreibungen

Die Modellierung von Paketwiederholungen infolge einer Zeitüberschreitung bedeutet, daß die Modelle eine Rückkopplungsschleife, die von der Durchlaufzeit der Anforderungen abhängt, enthalten müssen. Dies führt zu einer neuen Klasse von Warteschlangenmodellen. Bevor jedoch die abstrakten Verkehrsmodelle beschrieben werden, wird zuerst ein Modell eines Paketvermittlungsnetzes betrachtet.

9.3.1 Grundmodell für den Time-Out-Mechanismus

Als Basis für die Untersuchung dient das Verkehrsmodell nach Bild 9.4. Das zugrundeliegende Paketvermittlungsnetz ist modelliert als ein Warteschlangennetz, wobei die Warteschlangensysteme die Übertragungskanäle darstellen und die Verzweigungspunkte die wesentlich schnelleren Vermittlungsrechner repräsentieren. Insbesondere wird eine einzige Verkehrsbeziehung unter Berücksichtigung des Gesamtverkehrs im Netz untersucht. Dabei wird vorausgesetzt, daß all ihre Pakete über den gleichen Weg vermittelt werden.

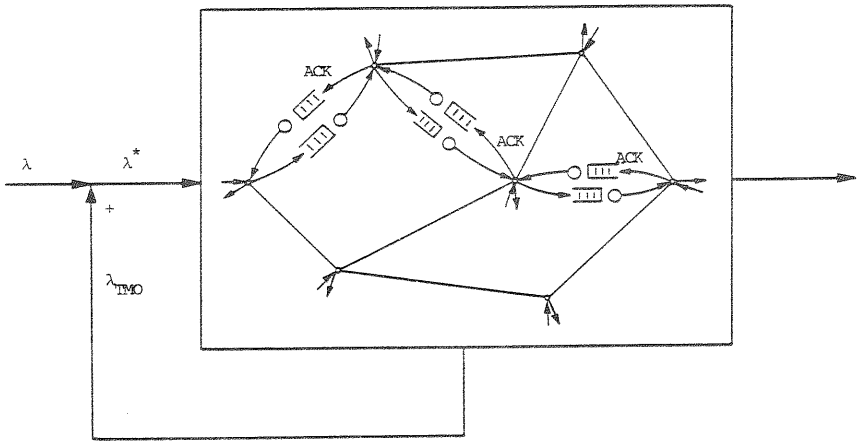


Bild 9.4: Basisverkehrsmodell: Time-Out Mechanismus.

In erster Linie ist hier der Einfluß von dem Time-Out-Mechanismus in einer der beiden höheren Ebenen (Vermittlungs- bzw. Transportebene) von Interesse. Deshalb wird angenommen, daß sämtliche Übertragungsfehler und eventuelle Speicherüberläufe von der Sicherungsebene aufgefangen werden. Darüberhinaus sollen keine Pakete verlorengehen durch bewußte Unterdrückung in einem überlasteten Netzknoten [Kamoun (1981)]. Pakete können daher nur noch beim Auftreten von Softwarefehlern oder durch Ausfall von Netzknoten verlorengehen. Unter normalen Umständen ist die Wahrscheinlichkeit für das Verschwinden von Paketen gering und wird deshalb im Modell vernachlässigt.

Unter den beschriebenen Voraussetzungen kann der Einfluß von einem zusätzlichen sogenannten Time-Out-Verkehr auf die Verkehrseigenschaften des Netzes untersucht werden. Dazu wird das nachfolgende Ablaufgeschehen betrachtet:
Pakete der untersuchten Verkehrsbeziehung treffen nach einer Poissonverteilung mit Rate λ ein. Sie werden durch das Netz zum Ziel vermittelt und die entsprechenden Quittungen (ACKnowledgements) werden zur Quelle zurückgeschickt. Eine Quittung wird

entweder vom Paketverkehr in der Gegenrichtung mitgeführt (Piggybacking) oder ist ein selbständiges Quittungspaket, wenn momentan keine Datenpakete in der Gegenrichtung vorhanden sind. Im Modell kann somit angenommen werden, daß jedes am Ziel ankommende Paket als Quittung zur Quelle zurückgesendet wird. Überschreitet die Quittierungszeit eines Paketes das eingestellte Time-Out-Intervall, so wird ein zusätzliches Paket (Kopie) abgeschickt. Alle Paketkopien treffen mit einer Gesamtrate λ_{TMO} ein, so daß der betrachteten Verkehrsbeziehung eine Rate $\lambda^* = \lambda + \lambda_{TMO}$ angeboten wird.

9.3.2 Analytisches Modell basierend auf einer Wahrscheinlichkeitsverteilung

a) Prinzipielles Berechnungsverfahren

Eine erste analytische Methode zur Berechnung der Time-Out-Rate λ_{TMO} ist in Bild 9.5 veranschaulicht. Im Mittelpunkt steht die komplementäre Verteilungsfunktion der Quittierungsverzögerung $F^C_{ACK}(t) = P\{T_{ACK} > t\}$, aus der, unter Berücksichtigung der zutreffenden Strategie, die mittlere Anzahl der Time-Outs berechnet werden kann.

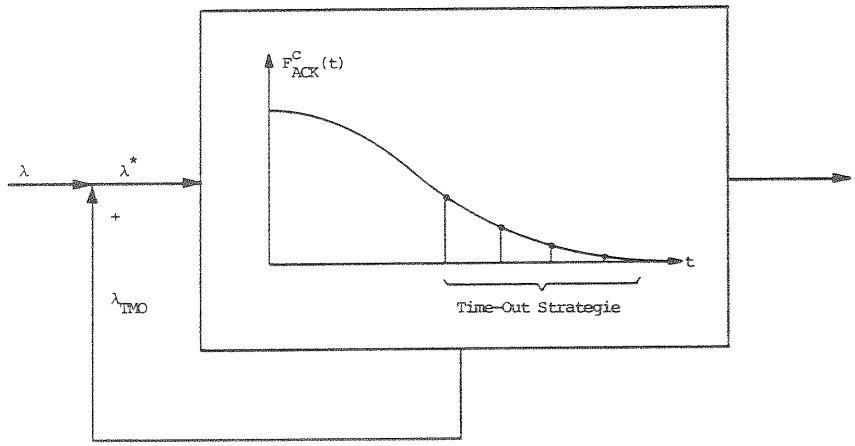


Bild 9.5: Verkehrsmodell: Time-Out Mechanismus (Berechnung einer Verteilungsfunktion der Quittierungsverzögerung).

Dazu betrachtet man die sequentiell eingestellten Zeitbegrenzungen T_i, $i = 1,2,...,$ eines Paketes und bestimmt die Wahrscheinlichkeit für die Überschreitung eines Time-Out-Zeitpunktes T_i:

$$F^C_{ACK}(T_1) = P\{T_{ACK} > T_1\} = P\{1\ TMO\} + P\{2\ TMO's\} + P\{3\ TMO's\} + ...$$

$$F^C_{ACK}(T_2) = P\{T_{ACK} > T_2\} = \quad P\{2\ TMO's\} + P\{3\ TMO's\} + ... \qquad (9.1)$$

$$F^C_{ACK}(T_3) = P\{T_{ACK} > T_3\} = \quad P\{3\ TMO's\} + ...$$

$$, \ T_1<T_2<T_3<.... \quad .$$

Wie aus diesem Schema hervorgeht, läßt sich die mittlere Anzahl der Paketwiederholungen $E[N_{TMO}]$ durch Summation sämtlicher Funktionswerte $F^C_{ACK}(T_i)$ ermitteln:

$$E[N_{TMO}] = \sum_{i\in z} F^C_{ACK}(T_i) = P\{1\ TMO\} + 2\cdot P\{2\ TMO's\} + 3\cdot P\{3\ TMO's\} + ... \qquad (9.2)$$

z : Menge aller Time-Out-Zeitpunkte der betrachteten Strategie

Als Spezialfall gilt für eine Time-Out Strategie mit einem konstanten Intervall T und mit einer unbegrenzten Wiederholmöglichkeit:

$$E[N_{TMO}] = \sum_{i=1}^{\infty} F^C_{ACK}(iT) \quad . \qquad (9.3)$$

Damit beträgt die Gesamtankunftsrate λ^*:

$$\lambda^* = \lambda\cdot(1+E[N_{TMO}]) = \lambda + \lambda_{TMO}. \qquad (9.4)$$

Bedingt durch die erhöhte Ankunftsrate λ^* ändert sich $F^C_{ACK}(t)$, so daß für den endgültigen Wert von λ_{TMO} eine iterative Prozedur erforderlich ist.

Dieses Berechnungsverfahren beruht auf zwei wesentlichen Annahmen:

- das Auftreten von Time-Outs ist unabhängig von der momentanen Netzbelastung, denn die Verteilungsfunktion kann lediglich eine gemittelte Netzsituation berücksichtigen,
- die Ankunftsabstände der Paketkopien sind ebenfalls negativ exponentiell verteilt, obwohl sie, wie Simulationen bestätigt haben, stärkere Streuungen aufweisen.

Aufgrund dieser Tatsachen ist das vorliegende Verfahren anwendbar, solange die Time-Out-Rate λ_{TMO} nur kleine Werte annimmt. Danach bricht das Netz zusammen, obwohl die analytische Berechnung noch für höhere Verkehrswerte fortgesetzt werden kann. Die Gültigkeit wurde durch Simulation verschiedener Verkehrsmodelle nachgeprüft (vgl. auch Bild 9.9). Die Anwendung umfaßt sowohl die offenen,als auch die geschlossenen Warteschlangennetze mit einer Produktlösungsform sowie Systeme, für die auf irgendeine Weise eine Quittierungszeitverteilungsfunktion ermittelt werden kann.

b) Offenes Warteschlangennetz

Für offene Netze, bestehend aus M/M/1-Warteschlangensystemen mit einer FIFO-Abfertigungsstrategie, kann gezeigt werden, daß die zufälligen Durchlaufzeiten in den einzelnen Warteschlangensystemen eines beliebigen, überholfreien Weges unabhängig voneinander und negativ exponentiell verteilt sind [Wong (1978), Walrand/Varaiya (1980), Melamed (1982), Daduna (1983)]. Die Verteilungsfunktion einer Summe von unabhängigen Zufallsvariablen ergibt sich aus der Faltung der einzelnen Verteilungsfunktionen. Die zugehörige Laplace-Stieltjes-Transformierte (LST) erfolgt demnach durch Produktbildung der einzelnen Laplace-Stieltjes-Transformierten.

Somit läßt sich für die LST der Quittierungszeitverteilungsfunktion $F_{ACK}(t)$ zwischen einem Ursprung A und einem Ziel B schreiben:

$$\Phi_{ACK}(s) = \prod_{i\in AB} \frac{1/t_{Fi}}{s+1/t_{Fi}} \cdot \prod_{j\in BA} \frac{1/t_{Fj}}{s+1/t_{Fj}} \quad . \qquad (9.5)$$

Hinweg AB Rückweg BA

Dabei ist t_{Fi} die mittlere Durchlaufzeit im M/M/1-Warteschlangensystem i [Kleinrock (1975)]:

$$t_{Fi} = \frac{h_i}{1-A_i} \quad , \tag{9.6}$$

wobei h_i die mittlere Bedienungszeit und $A_i = \lambda_i \cdot h_i$ das Verkehrsangebot des i-ten Systems im Hinweg AB ist. Die Ankunftsrate λ_i erhält man durch Lösung der linearen Gleichungen für das Verkehrsflußgleichgewicht (Gl.(4.43) mit k = 1). Hierbei ist zu berücksichtigen, daß dem Übertragungsweg zwischen Ursprung A und Ziel B ebenfalls die zusätzliche Time-Out-Rate λ_{TMO} angeboten wird.
Die Quittierungszeitverteilungsfunktion bekommt man durch Integration von $\frac{d}{dt} F_{ACK}(t)$ im Zeitbereich, d.h. durch Multiplikation von $\Phi_{ACK}(s)$ mit 1/s im Bildbereich und anschließender Rücktransformation:

$$F_{ACK}(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s} \cdot \Phi_{ACK}(s) \quad . \tag{9.7}$$

Und die entsprechende komplementäre Verteilungsfunktion ergibt sich aus:

$$F^{C}_{ACK}(t) = 1 - F_{ACK}(t) \quad . \tag{9.8}$$

Da $\Phi_{ACK}(s)$ eine rationale Funktion ist, läuft die Rücktransformation von Gl.(9.7) auf eine Partialbruchzerlegung folgender Gestalt hinaus [Doetsch (1976)]:

$$\Phi_{ACK}(s) = \frac{1}{s} + \sum_{k=1}^{K} \left[\frac{B_{1,k}}{(s+b_k)^{n_k}} + \frac{B_{2,k}}{(s+b_k)^{n_k-1}} + \ldots + \frac{B_{n_k-1,k}}{(s+b_k)^{2}} + \frac{B_{n_k,k}}{s+b_k} \right] + \sum_{m=1}^{M} \frac{C_m}{s+c_m} \tag{9.9}$$

Dabei gibt es:

- K Gruppen mit jeweils n_k Warteschlangensystemen mit der gleichen mittleren Durchlaufzeit: $b_k = 1/t_{Fi}$ bzw. $b_k = 1/t_{Fj}$ (mehrfache Pole); falls für Gruppe k alle Durchlaufzeiten in $\Phi_{ACK}(s)$ voneinander verschieden sind, entfallen diese Terme,
- M Warteschlangensysteme mit unterschiedlicher mittlerer Durchlaufzeit: $c_m = 1/t_{Fi}$ bzw. $c_m = 1/t_{Fj}$ (einfache Pole),

- insgesamt $1 + \sum_{k=1}^{K} n_k + M = 2N$ Terme; N = Anzahl der Übertragungsabschnitte AB.

In dieser allgemeinen Form lautet die Rücktransformation:

$$F_{ACK}(t) = 1 + \sum_{k=1}^{K} \left[\sum_{n=1}^{n_k} B_{n,k} \cdot \frac{t^{n-1}}{(n-1)!} \cdot e^{-b_k t} \right] + \sum_{m=1}^{M} c_m \cdot e^{-c_m t} \quad ; \quad t \geq 0 \tag{9.10}$$

Zur Bestimmung der Koeffizienten $B_{n,k}$, $n = 1,\ldots,n_k$, in der k-ten Gruppe definiert man die Funktion $\Psi_k(s)$:

$$\Psi_k(s) = (s+b_k)^{n_k} \cdot \Phi_{ACK}(s) \quad , \tag{9.11}$$

und erhält $B_{n,k}$ durch Auswertung der (n-1)-ten Ableitung von $\Psi_k(s)$ an der Stelle $s = -b_k$:

$$B_{n,k} = \frac{1}{(n-1)!} \cdot \Psi_k^{(n-1)}(s) \Big|_{s=-b_k} \quad . \tag{9.12}$$

Die notwendige Ableitung von $\Psi_k(s)$ erhält man rekursiv nach folgendem Schema, wobei Zähler $\alpha_k(s)$ und Nenner $\beta_k(s)$ der Funktion $\Psi_k(s)$ jeweils getrennt abgeleitet werden können:

$$\Psi_k(s) = \frac{1}{\beta_k(s)} \cdot [\alpha_k(s)]$$

$$\Psi_k^{(1)}(s) = \frac{1}{\beta_k(s)} \cdot [\alpha_k^{(1)}(s) - \beta_k^{(1)}(s) \cdot \Psi_k(s)]$$

$$\Psi_k^{(2)}(s) = \frac{1}{\beta_k(s)} \cdot [\alpha_k^{(2)}(s) - 2 \cdot \beta_k^{(1)}(s) \cdot \Psi_k^{(1)}(s) - \beta_k^{(2)}(s) \cdot \Psi_k(s)]$$

$$\Psi_k^{(3)}(s) = \frac{1}{\beta_k(s)} \cdot [\alpha_k^{(3)}(s) - 3 \cdot \beta_k^{(1)}(s) \cdot \Psi_k^{(2)}(s) - 3 \cdot \beta_k^{(2)}(s) \cdot \Psi_k^{(1)}(s) - \beta_k^{(3)}(s) \cdot \Psi_k(s)],$$

usw.

Allgemein gilt für die n-te Ableitung:

$$\Psi_k^{(n)}(s) = \frac{1}{\beta_k(s)} \cdot [\alpha_k^{(n)}(s) - \sum_{r=1}^{n} \binom{n}{r} \cdot \beta_k^{(r)}(s) \cdot \Psi_k^{(n-r)}(s)] \quad . \tag{9.13}$$

Entsprechend wird zur Ermittlung von C_m die Funktion $\Psi_m(s)$ definiert:

$$\Psi_m(s) = (s+c_m) \cdot \Phi_{ACK}(s) \quad , \tag{9.14}$$

mit der man C_m durch Auswertung an der Stelle $s = -c_m$ bekommt:

$$C_m = \Psi_m(s)\Big|_{s=-c_m} \quad . \tag{9.15}$$

Da bei Polynomen die Auswertung von Ableitungen mit dem Horner-Schema [Stiefel (1976)] numerisch durchgeführt werden kann, läßt sich die Rücktransformation von $\Phi_{ACK}(s)$ leicht automatisieren; vgl. auch [Olson/Brockus (1976)].

c) Geschlossene Warteschlangennetze

Auch für ein geschlossenes, überholfreies Warteschlangennetz läßt sich die LST der Quittierungszeit (Zykluszeit)-Verteilungsfunktion mathematisch angeben [Daduna (1982), Schassberger/Daduna (1983)]. Dabei werden die nicht betrachteten Netzverkehre und der Time-Out-Verkehr dadurch berücksichtigt, daß die Bedienungsrate in den einzelnen Warteschlangensystemen entsprechend reduziert werden (vgl. Abschnitt 4.3.2). Durch die hohe kombinatorische Zahl von Termen kann diese LST jedoch nicht mehr direkt numerisch behandelt werden. Als Ausweg steht die rekursive Berechnung der Momente der Quittierungszeitverteilungsfunktion nach [Reiser (1981b)] zur Verfügung. Mit Hilfe dieser Momente wird dann die Verteilungsfunktion approximiert, z.B. durch die Kombination einer exponentiellen Verteilung und einer Erlang-k-Verteilung [Kühn (1975), Jans (1983)].

9.3.3 Analytisches Modell basierend auf einer Zustandsbeschreibung

Durch die Verwendung einer Verteilungsfunktion kann der Time-Out-Mechanismus nur teilweise beschrieben werden, denn das dynamische Verhalten kann auf diese Weise nicht erfaßt werden. Für eine möglichst wirklichkeitsgetreue Modellierung des Time-Out-Effektes muß berücksichtigt werden, daß während einer kurzen Quittierungsverzögerung nur selten eine Zeitüberschreitung

stattfindet, währenddessen eine lange Verzögerung mit vielen Time-Out-Ereignissen verknüpft ist. Diese Zusammenhänge können in einem Modell, das auf einer Zustandsbeschreibung basiert, berücksichtigt werden.

Im folgenden wird ein Modellansatz für ein einstufiges Warteschlangensystem betrachtet. Dabei werden die Einschränkungen gemacht, daß jede Anforderung höchstens eine Kopie generieren kann und daß das Time-Out-Intervall konstant ist.

a) Modellansatz

In dem Modell (Bild 9.6) treffen Anforderungen aus einem Poisson Prozeß mit Rate λ im System ein. Abhängig von der zufälligen Anzahl der Anforderungen X, die sich im System befinden und die aus den ursprünglichen Anforderungen sowie deren Kopien bestehen, generiert eine ankommende Anforderung gleichzeitig ihre Kopie. Dieses geschieht, wenn X einen Grenzwert x_T überschreitet.

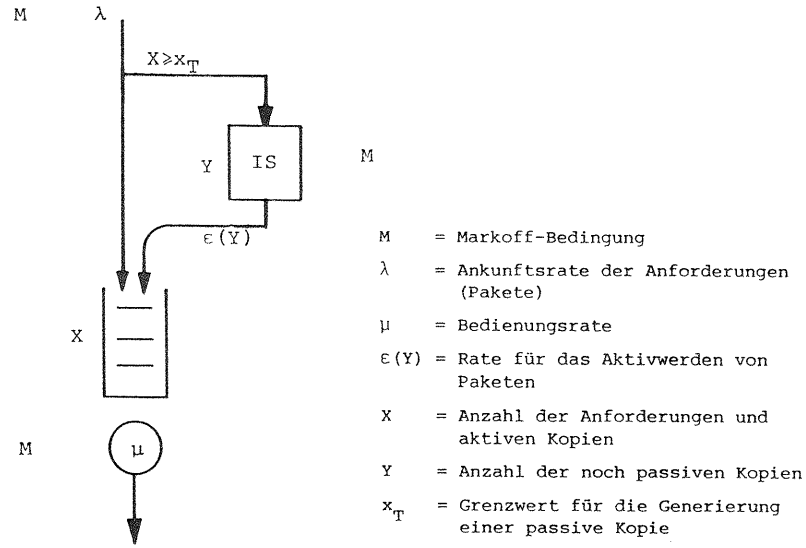
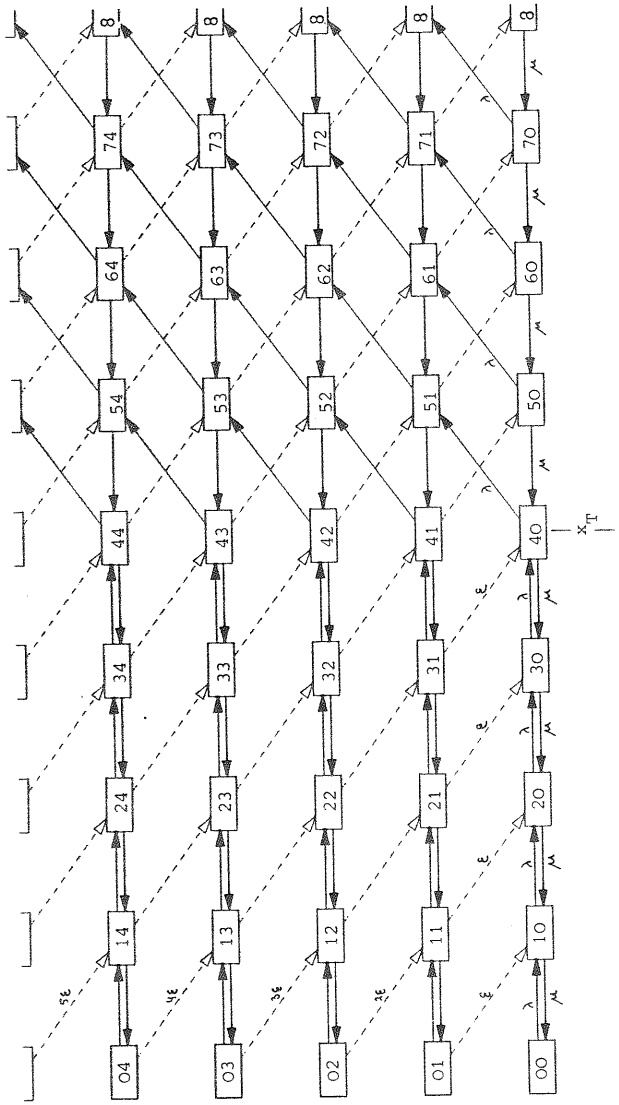


Bild 9.6: Verkehrsmodell: Time-Out Mechanismus (Zustandsbeschreibung).



Beispiel : $x_T = 4$

i j : Zustand (i,j)

i = Anzahl der Anforderungen und aktiven Kopien im Warteschlangensystem

j = Anzahl der noch passiven Kopien in der IS-Stufe

Bild 9.7: Zustandsdiagramm (Time-Out Mechanismus).

Diese Modellierung findet ihre Begründung darin, daß aus dem angetroffenen Systemzustand X bei der hier vorliegenden FIFO-Abfertigungsstrategie, eine gute Abschätzung für die erwartete Durchlaufzeit gemacht werden kann. Dadurch kann bereits beim Eintreffen im System entschieden werden, ob zusätzlich noch eine Kopie zu bedienen ist oder nicht. Bei einer exponentiell verteilten Bedienungszeit mit Mittelwert h und einem konstanten Time-Out-Intervall T ergibt sich für den Grenzwert x_T:

$$x_T = T/h \tag{9.16}$$

Die generierte Kopie wird in einem System mit unendlich vielen Bedienungseinheiten (Infinite-Server-Stufe) abgespeichert, das mit einer zustandsabhängigen Rate $\varepsilon(Y)$ verlassen wird. Beim Verlassen dieser IS-Stufe werden die Kopien aktiv und verursachen eine zusätzliche Systembelastung.

Der Zustandsprozeß läßt sich somit durch ein zweidimensionales Markoff-Zustandsdiagramm gemäß Bild 9.7 darstellen, wobei für den Zustand (X=i, Y=j) gilt:

i = Anzahl der Anforderungen und der aktiven Kopien im Warteschlangensystem, i = 0,...

j = Anzahl der noch passiven Kopien in der IS-Stufe, j = 0,...

Aus der Struktur des Zustandsdiagramms ist bereits der Time-Out-Mechanismus erkennbar. Ausgehend von einem leeren Systemzustand (0,0), verläuft der Zustandsprozeß zunächst wie in einem M/M/1-Warteschlangensystem. Wird jedoch der Grenzwert x_T überschritten, so bringt jede Anforderung eine zusätzliche Anforderung in der Form einer passiven Kopie mit. Dies drückt sich im Zustandsdiagramm durch einen Übergang in der aufsteigenden Diagonale aus: Zustand (i,j) nach Zustand (i+1,j+1). Bei jeder Abspeicherung einer passiven Kopie in der IS-Stufe nimmt die Rate $\varepsilon(j)$, mit der die Kopien aktiv werden, zu. Beim Verlassen der IS-Stufe findet eine Zustandsumspeicherung statt, so daß für die gestrichelt gezeichneten Zustandsübergänge (aktivwerdende Kopien) in der absteigenden Diagonale gilt:

Zustand (i,j) nach Zustand (i+1, j-1). Bei einem zu hohen Verkehrsangebot kann somit das Verkehrsmodell sich soweit "aufladen", daß es sich nicht mehr erholen kann: das System wird instabil.

9.3.4 Simulationsmodell

Zur Überprüfung der Genauigkeit des analytischen Modells und zur Untersuchung von adaptiven Strategien wurde der genaue Time-Out-Mechanismus für die in Bild 9.8 zusammengestellten Netzkonfigurationen simuliert.

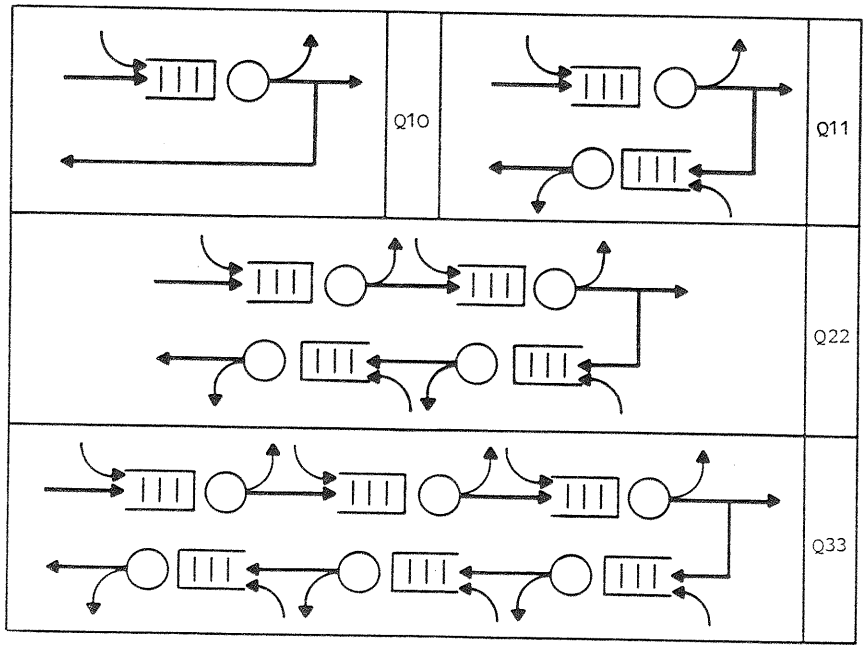


Bild 9.8: Simulationsmodell verschiedener Netzkonfigurationen zur Untersuchung des Time-Out Mechanismus.
Q10 : einstufiges Netz mit verzögerungsfreiem Rückkanal,
Q11 : einstufiges Netz in beide Richtungen,
Q22 : zweistufiges Netz in beide Richtungen,
Q33 : dreistufiges Netz in beide Richtungen.

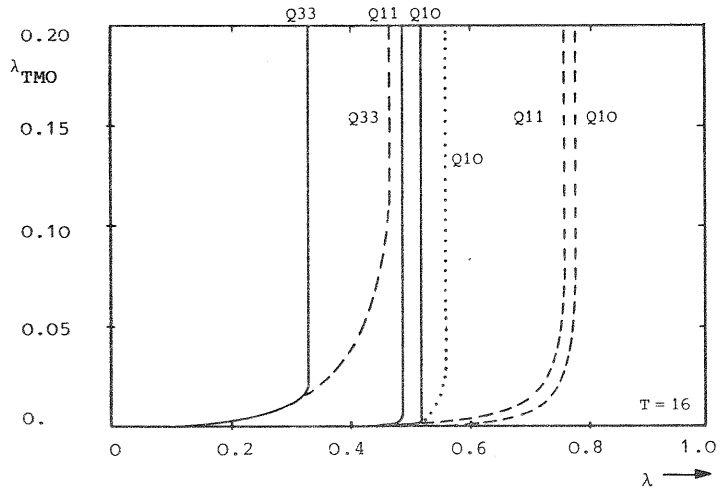
Es wird jeweils eine Ursprung - Ziel-Beziehung betrachtet, wobei weitere Verkehrsbeziehungen durch zusätzliche Netzverkehre, die nur ein Warteschlangensystem durchlaufen, berücksichtigt werden können. Alle Warteschlangensysteme haben eine unbegrenzte Kapazität, so daß keine Anforderungen abgewiesen oder blockiert werden. Anforderungen werden am Ziel direkt als Quittungen zurückgesendet. Kopien durchlaufen nur diejenigen Warteschlangensysteme, die zum Hinweg gehören.

9.4 Numerische Ergebnisse

Außer dem Vergleich zwischen den verschiedenen Verfahren werden nur simulative Ergebnisse gezeigt. Die Simulationen basieren auf jeweils 110.000 Anforderungen, deren Anzahl sich verteilt über einer Vorlaufphase und 10 Teiltests.Generierte Kopien und Anforderungen aus den zusätzlichen Netzquellen sind in dieser Zahl nicht enthalten. Die Resultate wurden im Abstand von $\Delta\lambda = 0.1$ mit 95% Vertrauensintervall ermittelt. Die Simulationspunkte und deren Vertrauensintervalle sind jedoch aus Übersichtlichkeitsgründen nicht eingetragen. Wo es aber für den genauen Kurvenverlauf erforderlich war, wurden genügend viele Zwischenwerte simuliert. Für alle gezeigten Kurven ist einheitlich eine mittlere Bedienungszeit $h = 1$ gewählt, und es wird keine zusätzliche Netzbelastung betrachtet. Untersuchungen mit Netzverkehr weisen jedoch ein ähnliches Verhalten auf [van As (1983)].

a) Vergleich der verschiedenen Verfahren

Zuerst werden die Untersuchungsverfahren miteinander verglichen. Insbesondere wird eine feste Einstellung der Zeitbegrenzung $T = 16$ betrachtet. In Bild 9.9 , das die Wiederholungsrate λ_{TMO} als Funktion der Ankunftsrate λ angibt, werden zwei Vergleiche durchgeführt. Zum einen werden analytische Ergebnisse, basierend auf einer Verteilungsfunktion für verschiedene Netzkonfigurationen mit Simulationsergebnissen verglichen, zum anderen werden die Kurven aller Verfahren für die Konfiguration Q10 einander gegenübergestellt.



Time-Out-Mechanismus: Wiederholungsrate λ_{TMO} als Funktion der angebotenen Rate λ.

Bild 9.9: Vergleich der verschiedenen Untersuchungsverfahren bei einem konstanten Time-Out-Intervall T = 16.
——— *Simulation*
— — — *Verteilungsfunktion*
····· *Zustandsbeschreibung*

Für den ersten Vergleich werden die gestrichelten Kurven, die mit einer Verteilungsfunktion nach Abschnitt 9.3.2 berechnet sind, mit den durchgezogenen Simulationskurven verglichen. Charakteristisch für die Resultate aller Netzkonfigurationen ist die abrupte und lawinenartige Zunahme von λ_{TMO}, währenddessen die instabilen Verkehrswerte, die aufgrund einer Verteilungsfunktion ermittelt werden, wesentlich höher liegen. Diese Diskrepanz wird verursacht durch den Mitkopplungseffekt des Time-Out-Mechanismus: wenn lange Quittierungsverzögerungen vorliegen, werden viele Kopien generiert und die dadurch erhöhte Netzbelastung verursacht ihrerseits eine noch größere Quittierungsverzögerung. Wie zu erwarten ist, nehmen die Unterschiede ab, wenn mehrere Warteschlangensysteme zwischen Ursprung und Ziel durchlaufen werden; denn in diesem Fall ist die Streuung der Quittierungsverzögerung kleiner. Ferner ist bemerkenswert, daß die Wiederholungsrate bei den Netzkonfigurationen Q10 und Q11 abrupt zunimmt, während

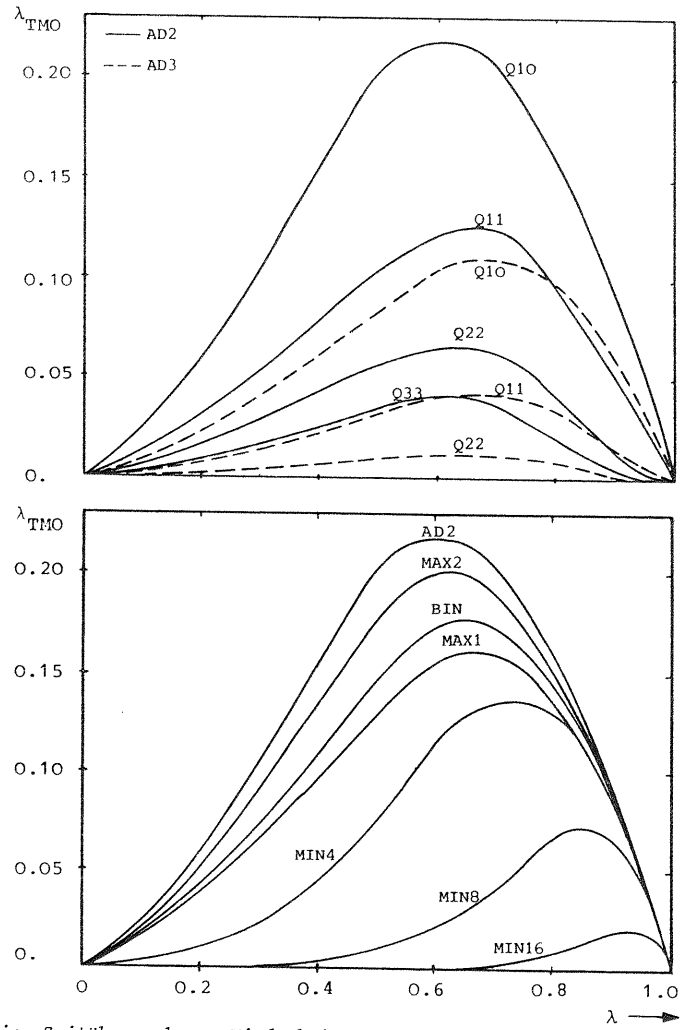
bei der Konfiguration Q33 die Time-Oute Rate λ_{TMO} am Anfang langsam steigt. In diesem Bereich decken sich die Resultate beider Verfahren.

Im zweiten Vergleich wird die Konfiguration Q10, bei der die größte Abweichung zwischen der analytischen und simulativen Methode festgestellt wurde, für sich betrachtet. Die punktierte Kurve gilt für das Verfahren, das auf einer Zustandsbeschreibung beruht . Obwohl die mögliche Anzahl von Kopien pro Anforderung auf eins beschränkt wird , kann der instabile Bereich bereits besser vorausgesagt werden. Die Voraussage ist umso besser, je größer der Zustandsraum gewählt wird, weil sich dadurch der Einfluß von Zustandsraumbegrenzungen verringert. Mann erkennt also, daß mit dem Modellansatz nach Abschnitt 9.3.3 die lawinenartige Zunahme von Kopien bei Überschreitung einer Grenzlast sehr gut beschrieben werden kann. Bei der Simulation wird das System etwa bei λ = 0.52 instabil. Aufgrund der punktierten Kurve liegt die Instabilität bei λ = 0.56.

b) Basisstrategie für die adaptive Zeitüberwachung

Bild 9.10 zeigt einige Simulationsergebnisse für den Verlauf der Wiederholungsrate λ_{TMO} als Funktion der angebotenen Rate λ, wenn die Zeitüberwachung adaptiv eingestellt wird. Das Bild enthält zwei Kurvenscharen für die vier verschiedenen Netzkonfigurationen.

Für die durchgezogene Kurvenschar AD2 (ADaptiv) wird die neue Einstellung des Time-Out-Intervalls doppelt so groß gewählt wie der Mittelwert der letzten 10 gemessenen Quittierungszeiten. Für die gestrichelte Kurvenschar AD3 wird der gemessene Mittelwert dreifach genommen.



Adaptive Zeitüberwachung: Wiederholungsrate λ_{TMO} als Funktion der angebotenen Rate λ.

Bild 9.10: Basisstrategie (Time-Out-Intervall proportional dem Mittelwert der letzten 10 Quittierungszeiten).
a) Verschiedene Proportionalitätsfaktoren: zweimal (AD2), dreimal (AD3).
b) Verschiedene Netzkonfigurationen.

Bild 9.11: Strategiemodifikationen, Netzkonfiguration Q10.
a) Vergleich mit Basisstrategie AD2.
b) Begrenzung der Anzahl Wiederholungen (MAX1, MAX2).
c) Binäre Verlängerung des nächsten Intervalls (BIN).
d) Minimalwert für das Time-Out Intervall (MIN4, MIN8, MIN16).

Der Verlauf der Kurven läßt verschiedene Folgerungen zu:

- Sowohl im niedrigen als auch im hohen Verkehrsbereich funktioniert die Anpassung an die momentane Belastungssituation im Netz sehr gut. Bei einem kleinen Verkehrsangebot ist die Quittierungszeit gering und nur in seltenen Fällen wird die Zeitbegrenzung überschritten. Bei einem hohen Verkehrsangebot ist ein ziemlich kontinuierlicher Verkehrsstrom vorhanden, der eine gute Schätzung für das nächste Time-Out-Intervall ermöglicht.

- In dem mittleren Verkehrsbereich jedoch erreicht die Wiederholungsrate λ_{TMO} teilweise wesentlich höhere Werte. Dies hängt mit folgender Tatsache zusammen. Trifft eine Gruppe von nacheinander dicht gestaffelten Anforderungen ein und ist gleichzeitig der Einstellungswert für das Time-Out-Intervall klein, dann wird, bis zur Korrektur dieses Wertes, die Zeitbegrenzung für einen großen Teil dieser Anforderungen zu klein eingestellt,und entsprechend viele Zeitbegrenzungen werden überschritten.

- Durch die Wahl eines größeren Multiplikators (AD3 anstatt AD2) wird der maximale Wert der Wiederholungsrate λ_{TMO} bedeutend geringer.

- Durch die geringere Streuung der Quittierungszeit wird bei einer größeren Anzahl von durchlaufenen Warteschlangensystemen (Übertragungsstrecken) das Netzverhalten unempfindlicher für den Time-Out-Mechanismus.

c) Strategiemodifikationen

Aus Bild 9.11 geht hervor, daß durch verschiedene Modifikationen die adaptive Einstellung der Zeitüberwachung noch weiter verbessert werden kann. Als Ausgangspunkt wird die Netzkonfiguration Q10 und die adaptive Strategie AD2 betrachtet.

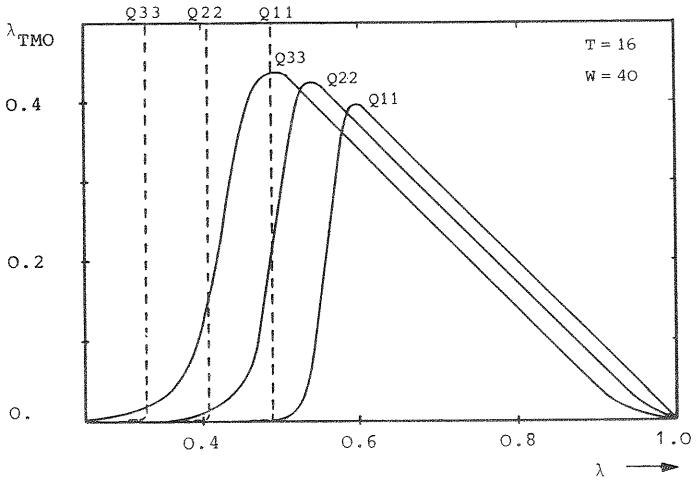
Die durch Paketwiederholungen verursachte Blindlast verringert sich nun durch:

- Begrenzung der Anzahl der Wiederholungen auf 2 (MAX2) oder 1 (MAX1).

- Binäre Verlängerung der Zeitbegrenzung für jede weitere Kopie eines Paketes: 2,4,8,16...mal des momentanen Einstellwertes (BIN).
- Keine Einstellung der Zeitbegrenzung unter einem vorgegebenen Zeitwert T = 4,8,16,... (MIN4, MIN8, MIN16).

d) Time-Out-Fenster

Bild 9.12 zeigt, daß mit Hilfe des Time-Out-Fensters sogar die vom Prinzip her instabile Strategie mit einer konstanten Einstellung der Zeitüberwachung unter Kontrolle gebracht werden kann. Als Time-Out-Fenster wurde W = 40 gewählt. Die Steigung der Kurven kann durch kleinere Wahl des Fensters noch beliebig abgeflacht werden. Zum Vergleich sind die Instabilitätsgrenzen der verschiedenen Netzkonfigurationen ohne diesen Mechanismus gestrichelt eingezeichnet.



Time-Out-Fenster: Wiederholungsrate λ_{TMO} als Funktion der angebotenen Rate λ.

Bild 9.12: Festes Time-Out-Intervall T = 16, Fenster W = 40.
---- ohne Time-Out-Fenster.
——— mit Time-Out-Fenster.

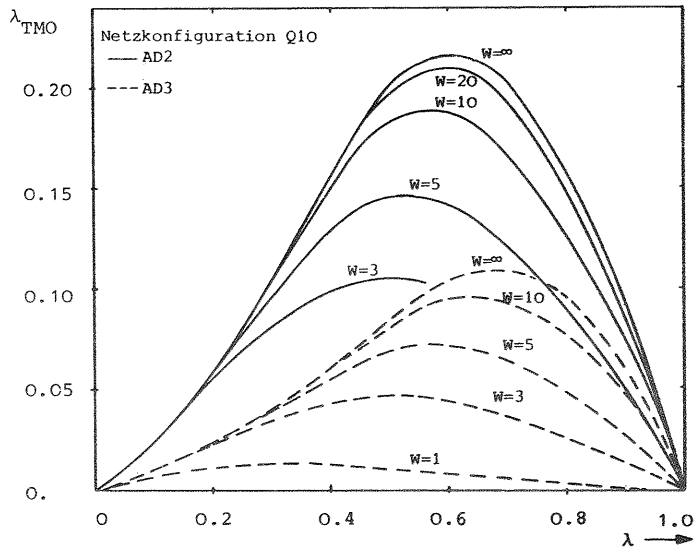


Bild 9.13: Adaptives Time-Out-Intervall und Time-Out-Fenster.
a) Verschiedene Proportionalitätsfaktoren: zweimal (AD2), dreimal (AD3).
b) Verschiedene Fenstergrößen (W = 1, 3, 5, 10, ∞).

Insbesondere in Zusammenhang mit adaptiven Strategien ist das Time-Out-Fenster eine zweckmäßige Ergänzung. Dies wird in Bild 9.13 verdeutlicht. Als Ausgangspunkt wird wieder die Netzkonfiguration Q10 betrachtet. Die beiden Referenzkurven sind AD2 und AD3 mit W = ∞.

9.5 Schlußfolgerung

Währenddessen der Time-Out-Mechanismus für eine gesicherte Ende-zu-Ende Datenübermittlung unentbehrlich ist, kann er in Überlastsituationen eine Lawine von wiederholten Paketen auslösen. Dadurch können die Verkehrseigenschaften des Netzes stark beeinträchtigt werden und die extrem hohe Blindlast kann sogar einen Zusammenbruch des Netzes verursachen.

In diesem Kapitel wurde jedoch gezeigt, daß durch Kombination einer adaptiven Einstellung der Zeitbegrenzung und eines Time-Out-Fensters das Problem der überflüssigen Paketwiederholungen völlig beseitigt werden kann.

10. ZUSAMMENFASSUNG

Die vorliegende Arbeit befaßte sich mit der Überlastproblematik in Paketvermittlungsnetzen.

Nach einem Überblick über typische Strukturen, Betriebsarten und Ablaufvorgängen in Paketvermittlungsnetzen wurden Ursachen und Indikatoren einer Überlastsituation sowie deren Überlastabwehrmaßnahmen diskutiert.

Zur quantitativen Erfassung einer Überlast und zur Leistungsbeurteilung der Überlastabwehrstrategien wurde die verkehrstheoretische Modellbildung herangezogen. Da das Entstehen einer Überlastsituation sowie die darauffolgenden Überlastabwehrmaßnahmen einen dynamischen Charakter haben, wurden verschiedene Verkehrsmodelle auch instationär betrachtet. Die dazu notwendigen mathematischen Hilfsmittel wurden bis zur programmtechnischen Implementierung bereitgestellt und erweitert. Um die verkehrstheoretischen Modelle effizient untersuchen zu können, wurden zwei Programmsysteme entwickelt: ein analytisches Programmsystem zur Durchführung der Routineaufgaben für die stationäre und instationäre Behandlung von mehrdimensionalen Markoff-Prozessen und ein leistungsfähiges, flexibles Simulationssystem zur Verkürzung der Programmerstellungszeit für ein stochastisches Simulationsmodell.

Das Zustandekommen von charakteristischen Überlastsituationen wurde statistisch beschrieben und quantitativ erfaßt. Es handelte sich hierbei um die Ausbreitung von Überlastsituationen im Netz, den Rückstau von Paketen, die lawinenartige Überflutung des Netzes mit Paketkopien und den Einfluß einer begrenzten Speicherkapazität.

In den anschließenden Kapiteln wurden analytische und simulative Verkehrsmodelle für die Untersuchung der Grundmechanismen von Überlastabwehrstrategien entwickelt und analysiert. Betrachtet wurden die Regelung der Netzzugänge, die Auslagerung, die Zweipunkt-Regelung und die adaptive Zeitbegrenzung. Ein wesentliches Ziel hierbei war, die Dynamik der einzelnen Überlastabwehrstrategien zu erfassen. Zur eindeutigen Identifizierung der

Systemreaktionen wurden rechteckige Überlastimpulse verwendet. Die Methodik läßt jedoch auch beliebige Impulsformen zu.

Die in dieser Arbeit gewonnenen Erkenntnisse sollen dazu beitragen, die einzelnen Überlastabwehrmaßnahmen in einem hierarchischen Überlaststeuerungsplan derart aufeinander abzustimmen, daß Stausituationen in heutigen und zukünftigen Paketvermittlungsnetzen weitgehend verhindert werden können.

LITERATURVERZEICHNIS

A. O. Allen: Probability, Statistics and Queueing Theory, Academic Press, New York/San Francisco/London, (1978).

H. R. van As: Simulation Models for Computer Controlled Telephone Exchanges, Intern. Symposium SIMULATION '75, Zurich, (1975),S.282-287, Acta Press (Ed. M. H. Hamza), Calgary/Zurich.

H. R. van As: Congestion Control in Gateway Nodes of Local Networks, NTG/GI-Fachtagung Struktur und Betrieb von Rechensystemen, Ulm (1982), S. 192-203, NTG-Fachberichte Band 80, VDE-Verlag GmbH, Berlin/Offenbach.

H. R. van As: Time-Out Management in Packet Switching Networks, 10th Intern. Teletraffic Congress, Montreal (1983), Congressbook, paper 3.3.2.

H. R. van As: Congestion Control in Packet Switching Networks by a Dynamic Foreground-Background Storage Strategy, 2nd Intern. Symposium on the Performance of Computer-Communication Systems, Zurich (1984), S. 433-448, North-Holland Publishing Company (Ed. H. Rudin, W. Bux), Amsterdam/New York/Oxford.

H. R. van As: QSIMLIB, A Modular FORTRAN IV Package Supporting Development of Stationary and Transient Simulation Programs for Complex Queueing Systems, Interner Bericht, FG Nachrichtentechnik, Universität Siegen (1984).

F. Baskett, K. M. Chandy, R. R. Muntz, F. G. Palacios: Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, J. Assoc. Comput. Mach. 29 (1975), S. 248-260.

R. R. Boorstyn, A. Livne : A Technique for Adaptive Routing in Networks, IEEE Trans. Commun. COM-29 (1981), S. 474-480.

P. J. Burke: The Output of a Queuing System, Operations Research 4 (1956), S. 699-704.

M. Butto, G. Colombo, G. Taggiasco, A. Tonietti: Models for the Performance Evaluation of a Packet Switching Network with Retransmission Time-Out, National Telecommunication Conference, Los Angeles (1977), S. 12.3.1/6.

W. Bux: Modelling and Analysis of Store-and-Forward Data Switching Centres with Finite Buffer Memory and Acknowledgement Signalling, 8th Intern. Teletraffic Congress, Melbourne (1976), Congressbook, paper 524.

W. Bux: Single Server Queues with General Interarrival and Phase-Type Service Time Distributions - Computational Algorithms, 9th Intern. Teletraffic Congress, Torremolinos (1979), Congressbook, paper 413.

W. Bux, U. Herzog: Approximation von Verteilungsfunktionen, ein wichtiger Schritt bei der Modellbildung für Rechensysteme, Workshop über Modelle für Rechensysteme, Bonn (1977), Informatik-Fachberichte Band 9, Springer Verlag, Berlin/Heidelberg/New York,S.1-15.

W. Bux, U. Herzog: The Phase-Concept: Approximation of Measured Data and Performance Analysis, Intern. Symposium on Computer Performance Modeling, Measurement, and Evaluation, Yorktown Heights, NY (1977), S. 23-28, North-Holland Publishing Company (Ed. K. M. Chandy, M. Reiser), Amsterdam/New York/Oxford.

W. Bux, P. J. Kühn, K. Kümmerle: Throughput Considerations in a Multi-Processor Packet Switching Node, IEEE Trans. Commun. COM-27 (1979), S. 745-750.

J. P. Buzen: Computational Algorithms for Closed Queueing Networks with Exponential Servers, Commun. ACM 16 (1973), S. 527-531.

K. M. Chandy, D. Neuse: Linearizer: A Heuristic Algorithm for Queueing Network Models of Computing Systems, Commun. ACM 25 (1982), S. 126-141.

K. M. Chandy, C. H. Sauer: Computational Algorithms for Product Form Queueing Networks, Commun. ACM 23 (1980), S. 573-583.

A. Chatterjee, N. D. Georganas, P. K. Verma: Analysis of a Packet-Switched Network with End-to-End Congestion Control and Random Routing, IEEE Trans. Commun. COM-25 (1977), S. 1485-1489.

W. Chou, A. W. Bragg, A. A. Nilsson: The Need for Adaptive Routing in the Chaotic and Unbalanced Traffic Environment, IEEE Trans. Commun. COM-29 (1981), S. 481-490.

W. W. Chu, G. Fayolle, D. Hibbits: An Analysis of a Tandem Queueing System for Flow Control in Computer Networks, IEEE Trans. Computers C-30 (1981), S. 318-324.

W. W. Chu, M. Shen: A Hierarchical Routing and Flow Control Policy for Packet Switched Networks, IEEE Trans. Computers C-29 (1980), S. 971-977.

E. Çinlar: Introduction to Stochastic Processes, Prentice Hall, Inc., Englewood Cliffs, New Jersey (1975).

J. W. Cohen: The Single Server Queue, North-Holland Publishing Company, Amsterdam/London (1969), 2. Auflage (1982).

R. B. Cooper: Introduction to Queueing Theory, 1. Auflage: The Macmillan Company, New York/Collier-Macmillan Limited, London (1972), 2. Auflage: Elsevier North Holland, New York/Oxford (1981).

P. J. Courtois: Decomposability, Queueing and Computer System Applications, Academic Press, New York/San Francisco/London (1977),

D. R. Cox, H. D. Miller: The Theory of Stochastic Processes, Methuen & Co., London (1965).

H. Daduna: Passage Times for Overtake-Free Paths in Gordon-Newell Networks, Adv. Appl. Prob. 14 (1982), S. 672-686.

H. Daduna: On Passage Times in Jackson Networks: Two-Stations Walk and Overtake-Free Paths, Zeitschrift für Operations Research 27 (1983), S. 239-256.

D. W. Davies: The Control of Congestion in Packet-Switching Networks, IEEE Trans. Commun. COM-20 (1972), S. 546-550.

D. W. Davies, D. L. A. Barber, W. L. Price, C. M. Solomonides: Computer Networks and Their Protocols, John Wiley & Sons, Chichester/New York/Bisbane/Toronto (1979).

G. Dehl: Prozeduren zur modularisierten Simulation, Diplomarbeit Nr. 16, FG Nachrichtentechnik, Universität Siegen (1981).

W. Dieterle: A Simulation Study of CCITT X.25: Throughput Classes and Window Flow Control, 10th Intern. Teletraffic Congress, Montreal (1983), Congressbook, paper 3.3.6.

G. Dietrich, R. Salade: Subcall-Type Control Simulation of SPC Switching Systems, Electrical Communication 52 (1977), S. 54-61.

G. Doetsch: Einführung in Theorie und Anwendung der Laplace-Transformation, Birkhäuser-Verlag, Basel/Stuttgart, 3. Auflage (1976).

DP-ISO-Basic Reference Model: Data Processing - Open Systems Interconnection-Basic Reference Model, Computer Networks 5 (1981), S. 81-118.

Ch. Z. Druckarch, J. M. van den Burg: The Dutch Datanet-General Aspects and Typical Network Structure, Networks 80 Intern. Conf. on Data Networks, London (1980), S. 233-249.

G. Fayolle, E. Gelenbe, G. Pujolle: An Analytic Evaluation of the Performance of the 'Send and Wait' Protocol, IEEE Trans. Commun. COM-26 (1978), S. 313-319.

W. Feller: An Introduction to Probability Theory and its Applications, Vol. I, John Wiley & Sons, New York/London/Sydney (3rd Ed. 1968).

W. Feller: An Introduction to Probability Theory and its Applications, Vol. II, John Wiley & Sons, New York/London/Sydney (1966).

F. Finger: Interprozessor-Kommunikation über ein Koppelnetz in Multirechnersystemen - Untersuchungen zum Verkehrsverhalten, Diplomarbeit Nr. 34, FG Nachrichtentechnik, Universität Siegen (1983).

W. Fischer: Untersuchung des verallgemeinerten geschalteten Poisson-Prozesses zur Beschreibung des Steueraufrufverkehrs in rechnergesteuerten Vermittlungssystemen, Studienarbeit Nr. 35, FG Nachrichtentechnik, Universität Siegen (1983).

G. S. Fishman: Concepts and Methods in Discrete Event Digital Simulation, John Wiley & Sons, New York/London/Sydney/Toronto (1973).

G. S. Fishman: Principles of Discrete Event Simulation, John Wiley & Sons, New York/Brisbane/Chichester/Toronto (1978).

H. C. Folts: X.25 Transaction-Oriented Features - Datagram and Fast Select, IEEE Trans. Commun. COM-28 (1980), S. 496-500.

H. Gabler, W. Tietz: Datenkommunikation in den Fernmeldenetzen der Deutschen Bundespost, Informatik-Spektrum 4 (1981), S. 11-30.

R. Garcia: Eine Planungsmethode für das Vermittlungsnetz eines Datennetzes mit Paketvermittlung, OR Spektrum 4 (1982), S.237-244.

E. Gelenbe, I. Mitrani: Analysis and Synthesis of Computer Systems, Academic Press, London/New York/Toronto/Sydney/San Francisco (1980).

N. D. Georganas: Modeling and Analysis of Message Switched Computer-Communication Networks with Multilevel Flow Control, Computer Networks 4 (1980), S. 285-294.

M. Gerla, L. Kleinrock: Flow Control: A Comparative Survey, IEEE Trans. Commun. COM-28 (1980), S. 553-574.

A. Giessler, J. Hänle, A. König, E. Pade: Free Buffer Allocation - An Investigation by Simulation, Computer Network 2 (1978), S. 191-208.

A. Giessler, A. Jägemann , E. Mäser, J. O. Hänle: Flow Control Based on Buffer Classes, IEEE Trans. Commun. COM-29 (1981), S. 436-443.

A. Giessler, A. Jägemann, E. Mäser: Simulation of an X.25 Network providing Throughput Guaranties, Intern. Conf. on Performance of Data Communication Systems and their Applications, Paris (1981), S. 279-290, North-Holland Publishing Company (Ed. G. Pujolle), Amsterdam/New York/Oxford.

L. A. Gimpelson: Network Management: Design and Control of Communications Networks, Electrical Communication 49 (1974), S. 4-22.

G. Gordon: System Simulation, Prentice Hall, Englewood Cliffs, N.J., (2nd Ed. 1978).

W. H. Greene, U. W. Pooch: A Review of Classification Schemes for Computer Communication Networks, Computer 10 (1977),S.12-21.

D. Gross, C. M. Harris: Fundamentals of Queueing Theory, John Wiley & Sons, New York/London/Sydney/Toronto (1974).

K. D. Günther: Prevention of Deadlocks in Packet-Switched Data Transport Systems, IEEE Trans. Commun. COM-29 (1981),S.512-524.

D. G. Haenschke, D. A. Kettler, E. Oberer: Network Management and Congestion in the U.S. Telecommunications Network, IEEE Trans. Commun. COM-29 (1981), S. 376-385.

P. G. Harrison: Performance Prediction of a Flow Control System Using an Analytical Model, Intern. Symposium on Local Computer Networks, Florence (1982), S. 439-458, North Holland Publishing Company (Ed. P. C. Ravasio, G. Hopkins, N. Naffah), Amsterdam/ New York/Oxford.

U. Herzog, L. Woo, K. M. Chandy: Solution of Queueing Problems by a Recursive Technique, IBM J. Res. Develop. 19 (1975), S. 295-300.

D. P. Heyman, M. J. Sobel: Stochastic Models in Operations Research, Vol. I: Stochastic Processes and Operating Characteristics, McGraw-Hill Book Company, New York (1982).

F. Hillebrand: DATEX, Infrastruktur der Daten- und Textkommunikation, R. v. Decker's Verlag, G. Schenck, Heidelberg/Hamburg (1981).

M. H. van Hoorn: Algorithms for the State Probabilities in a General Class of Single Server Queueing Systems with Group Arrivals, Management Science 27 (1981), S. 1178-1187.

M. H. van Hoorn: Algorithms and Approximations for Queueing Systems, Dissertation, Vrije Universiteit Amsterdam (1983). auch erschienen als Mathematical Centre Tract, Mathematisch Centrum, Amsterdam (1984).

D. L. Iglehart, G. S. Shedler: Regenerative Simulation of Response Times in Networks of Queues, Lecture Notes in Control and Information Sciences, Band 26, Springer-Verlag, Berlin/ Heidelberg/New York (1980).

M. I. Irland: Buffer Management in a Packet Switch, IEEE Trans. Commun. COM-26 (1978), S. 328-337.

M. I. Irland, G. Pujolle: Comparison of Two Packet Retransmission Techniques, IEEE Trans. Inform. Theory IT-26 (1980), S. 92-97.

H. Jans: Verkehrsanalyse von Vermittlungs-Steuerungen mit taktgesteuerter Ein-/Ausgabe und Prioritäten, Dissertation, Universität Siegen (1983).

F. Kamoun: Design Considerations for Large Computer Communications Networks, Ph. D. Dissertation, University of California, Los Angeles (1976).

F. Kamoun: A Drop and Throttle Flow Control Policy for Computer Networks, IEEE Trans. Commun. COM-29 (1981), S. 444-452.

F. Kamoun, L. Kleinrock: Analysis of Shared Finite Storage in a Computer Network Node Environment Under General Traffic Conditions, IEEE Trans. Commun. COM-28 (1980), S. 992-1003.

G. Kampe: SIMSCRIPT, Vieweg-Verlag, Braunschweig (1971).

J. S. Kaufman: Blocking in a Shared Resource Environment, IEEE Trans. Commun. COM-29 (1981), S. 1471-1481.

L. Kaufman, B. Gopinath, E. F. Wunderlich: Analysis of Packet Network Congestion Control Using Sparse Matrix Algorithms, IEEE Trans. Commun. COM-29 (1981), S. 453-465.

P. Kermani: Analysis of a Feedback Scheme for Congestion Control im Computer Networks, Intern. Conf. on Performance of Data Communication Systems and their Applications, Paris (1981), S. 331-343, North-Holland Publishing Company (Ed. G. Pujolle), Amsterdam/New York/Oxford.

P. Kermani, L. Kleinrock: Analysis of Buffer Allocation Schemes in a Multiplexing Node, Intern. Conference on Communications, Chicago (1977), S. 30.4-269/275.

P. Kermani, L. Kleinrock: Dynamic Flow Control in Store-and-Forward Computer Networks, IEEE Trans. Commun. COM-28 (1980), S. 263-271.

H. Kerner, G. Bruckner: Rechnernetzwerke: Systeme, Protokolle und das ISO-Architekturmodell, Springer-Verlag, Wien/New York (1981).

W. M. Kiesel, P. J. Kühn: A New Multi-Access Protocol with Dynamic Priorities for Distributed Systems, NTG/GI-Fachtagung Struktur und Betrieb von Rechensystemen, Ulm (1982), S. 181-191, NTG-Fachberichte Band 80, VDE-Verlag GmbH, Berlin/Offenbach.

L. Kleinrock: Communication Nets, Stochastic Message Flow and Delay, McGraw-Hill Book Company, New York/San Francisco/Toronto, London (1964).

L. Kleinrock: Queueing Systems, Vol. I: Theory, John Wiley & Sons, New York/London/Sydney/Toronto (1975).

L. Kleinrock: Queueing Systems, Vol. II: Computer Applications, John Wiley & Sons, New York/London/Sydney/Toronto (1976).

L. Kleinrock, P. Kermani: Static Flow Control in Store-and - Forward Computer Networks, IEEE Trans. Commun. COM-28 (1980), S. 271-279.

L. Kleinrock, C. W. Tseng: Flow Control Based on Limiting Permit Generation Rates, 5th Intern. Conf. on Computer Communication, Atlanta (1980), S. 785-790.

U. Körner: Congestion Control in Packet Switching Computer Communication Networks, 10th Intern. Teletraffic Congress, Montreal (1983), Congressbook, paper 3.3.5.

L. Kosten: Simulation in Traffic Theory, 6th Intern. Teletraffic Congress, Munich (1970), Congressbook, paper 411.

P. J. Kühn: Über die Berechnung der Wartezeiten in Vermittlungs- und Rechnersystemen, Dissertation, Universität Stuttgart (1972).

P. J. Kühn: Zur optimalen Steuerung des Multiprogramminggrades in Rechnersystemen mit virtuellen Speicher und Paging, GI-5. Jahrestagung, Dortmund (1975), S. 567-580, Lecture Notes in Computer Science, Bd. 34, Springer-Verlag, Berlin/Heidelberg/New York.

P. J. Kühn: Approximate Analysis of General Queuing Networks by Decomposition, IEEE Trans. Commun. COM-27 (1979), S. 113-126.

P. J. Kühn: Analyse zufallsabhängiger Prozesse in Systemen zur Nachrichtenvermittlung und Nachrichtenverarbeitung, Habilitation, Universität Stuttgart (1981).

P. J. Kühn: Analysis of Busy Periods and Response Times in Queueing Networks by the Method of First Passage Times, 9th International Symposium on Computer Performance Modelling, Measurement, and Evaluation, College Park, Maryland, USA (1983), S. 437-456, North-Holland Publishing Company (Ed. A. K. Agrawala, S. K. Tripathi), Amsterdam/New York/Oxford.

J. Labetoulle, G. Pujolle: A Study of Flows Through Virtual Circuits Computer Networks, Computer Networks 5 (1981), S. 119-126.

S. S. Lam: Store-and-Forward Buffer Requirements in a Packet Switching Network, IEEE Trans. Commun. COM-24 (1976), S. 394-403.

S. S. Lam: Queuing Networks with Population Size Constraints, IBM J. Res. Develop. 21 (1977), S. 370-378.

S. S. Lam, Y. L. Lien: Congestion Control of Packet Communication Networks by Input Buffer Limits - A Simulation Study, IEEE Trans. Computers C-30 (1981), S. 733-742.

S. S. Lam, Y. L. Lien: A Tree Convolution Algorithm for the Solution of Queueing Networks, Commun. ACM 26 (1983), S. 203-215.

S. S. Lam, M. Reiser: Congestion Control of Store-and-Forward Networks by Input Buffer Limits - An Analysis, IEEE Trans. Commun. COM-27 (1979), S. 127-134.

S. S. Lam, J. W. Wong: Queueing Network Models of Packet Switching Networks, Part 2: Networks with Population Size Constraints, Performance Evaluation 2 (1982). S. 161-180.

G. Lamprecht: Einführung in die Programmiersprache SIMULA, Vieweg Verlag, Braunschweig (1976).

G. Latouche: Exponential Servers Sharing a Finite Storage: Comparison of Space Allocation Policies, IEEE Trans. Commun. COM-28 (1980), S. 910-915.

S. S. Lavenberg: Computer Performance Modeling Handbook, Academic Press, New York/London/Paris/San Diego/San Francisco/Sao Paulo/Sydney/Tokyo/Toronto (1983).

A. M. Law, W. D. Kelton: Simulation Modeling and Analysis, McGraw-Hill Book Company, New York (1982).

C. Lemieux: Theory of Flow Control in Shared Networks and Its Application in the Canadian Telephone Network, IEEE Trans. Commun. COM-29 (1981), S. 399-413.

J. D. C. Little: A Proof of the Queueing Formula L=λW, Operations Research 9 (1961), S. 383-387.

J. C. Majithia, M. Irland, J. L. Grangé, N. Cohen, C. O'Donnell: Experiments in Congestion Control Techniques, Intern. Symposium on Flow Control in Computer Networks, Versailles (1979), S. 211-234, North-Holland Publishing Company (Ed. J. L. Grangé, M. Gien), Amsterdam/New York/Oxford.

J. Majus: Deadline Scheduling in Packet Switching Systems with Delay - Advantages and Constraints - , Intern. Conf. on Performance of Data Communication Systems and their Applications, Paris (1981), S. 399-408, North-Holland Publishing Company (Ed. G. Pujolle), Amsterdam/New York/Oxford.

D. Manfield, P. Tran-Gia: Queueing Analysis of an Arrival-Driven Message Transfer Protocol, 10th Intern. Teletraffic Congress, Montreal (1983), Congressbook, paper 4.1.4.

M. J. Maron: Numerical Analysis: A Practical Approach, Macmillan Publishing Company, New York; Collier Macmillan Publishers, London (1982).

J. Matsumoto, H. Mori: Flow Control in Packet-Switched Networks by Gradual Restrictions of Virtual Calls, IEEE Trans. Commun. COM-29 (1981), S. 466-473.

J. M. McQuillan: Interactions between Routing and Congestion Control in Computer Networks, Intern. Symposium on Flow Control in Computer Networks, Versailles (1979), S. 63-75, North-Holland Publishing Company (Ed. J. L. Grangé, M. Gien), Amsterdam/New York/Oxford.

J. Mehdi: Stochastic Processes, Wiley Eastern Limited, New Delhi/Bangalore/Bombay/Calcutta (1981).

B. Melamed: Sojourn Times in Queueing Networks, Mathematics of Operations Research 7 (1982), S. 223-244.

P. M. Merlin: Specification and Validation of Protocols, IEEE Trans. Commun. COM-27 (1979), S. 1671-1680.

P. M. Merlin, P. J. Schweitzer: Deadlock Avoidance in Store-and-Forward Networks, I: Store-and-Forward Deadlock, II: Other Deadlock Types, IEEE Trans. Commun. COM-28 (1980), S. 345-360.

R. J. T. Morris: Fixing Timeout Intervals for Lost Packet Detection in Computer Communication Networks, National Computer Conference 48 (1979), S. 887-891.

B. Müller: Zerlegungsorientierte, numerische Verfahren für Markovsche Rechensystemmodelle, Dissertation, Universität Dortmund (1980).

A. R. Odoni, E. Roth: An Empirical Investigation of the Transient Behavior of Stationary Queueing Systems, Operations Research 31 (1983), S. 432-455.

C. F. Olsen, C. G. Brockus: Digital Computation of Inverse Laplace Transforms, Simulation 27 (1976), S. 197-202.

M. Pennotti, M. Schwartz: Congestion Control in Store-and-Forward Tandem Links, IEEE Trans. Commun. COM-23 (1975), S. 1434-1443.

L. Pouzin: Methods, Tools, and Observations on Flow Control in Packet-Switched Data Networks, IEEE Trans. Commun. COM-29 (1981), S. 413-426.

W. L. Price: Data Network Simulation Experiments at the National Physical Laboratory, 1968-76, Computer Networks 1 (1977), S. 199-210.

G. Pujolle: The Influence of Protocols on the Stability Condition in Packet Switching Networks, IEEE Trans. Commun. COM-27 (1979), S. 611-619.

G. Pujolle: Comparison of Some End-to-End Flow Control Policies in a Packet-Switching Network, National Computer Conference 48 (1979), S. 893-903.

E. Raubold, J. Hänle: A Method of Deadlock-Free Resource Allocation and Flow Control in Packet Networks, Intern. Conference on Computer Communications, Toronto (1976), S. 483-487.

M. Reiser: Numerical Methods in Separable Queueing Networks, TIMS Studies in the Management Sciences 7 (1977), S. 113-142.

M. Reiser: A Queueing Network Analysis of Computer Communication Networks with Window Flow Control, IEEE Trans. Commun. COM-27 (1979), S. 1199-1209.

M. Reiser: Mean Value Analysis and Convolution Method for Queue-Dependent Servers in Closed Queueing Networks, Performance Evaluation 1 (1981), S. 7-18.

M. Reiser: Calculation of Response-Time Distributions in Cycle Exponential Queues, Performance Evaluation 1 (1981), S. 331-333.

M. Reiser: Admission Delays on Virtual Routes with Window Flow Control, Intern. Conf. on Performance of Data Communication Systems and their Applications, Paris (1981), S. 67-76, North-Holland Publishing Company (Ed. G. Pujolle), Amsterdam/ New York/Oxford.

M. Reiser: Performance Evaluation of Data Communications Systems, Proc. IEEE 70 (1982), S. 171-196.

L. G. Roberts: The Evolution of Packet Switching, Proc. IEEE 66 (1978), S. 1307-1313.

H. Rohlfing: SIMULA, Eine Einführung, Hochschultaschenbücher Band 747, Bibliographisches Institut, Mannheim/Wien/Zürich(1973).

H. Rösmann: Simulation mit GPSS, R. Oldenburg Verlag, München/ Wien (1978).

S. M. Ross: Stochastic Processes, John Wiley & Sons, New York/ Chichester/Brisbane/Toronto/Singapore (1983).

H. Rudin: On Routing and Delta Routing: A Taxonomy and Performance Comparison of Techniques for Packet-Switched Networks, IEEE Trans. Commun. COM-24 (1976), S. 43-59.

H. Rudin, H. Mueller: Dynamic Routing and Flow Control, IEEE Trans. Commun. COM-28 (1980), S. 1030-1039.

H. Rutishauser: Vorlesungen über numerische Mathematik, Band 1: Gleichungssysteme, Interpolation und Approximation, Birkhäuser Verlag, Basel/Stuttgart (1976).

H. Rutishauser: Vorlesungen über numerische Mathematik, Band 2: Differentialgleichungen und Eigenwertprobleme, Birkhäuser Verlag, Basel/Stuttgart (1976).

A. Rybczynski: X.25 Interface and End-to-End Virutal Circuit Service Characteristics, IEEE Trans. Commun. COM-28 (1980), S. 500-510.

S. Saad, M. Schwartz: Input Buffer Limiting Mechanisms for Congestion Control, Intern. Conference on Communications, Seattle (1980), S. 23.1.1.-5.

C. H. Sauer, K. M. Chandy: Computer Systems Performance Modeling, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1981).

U. Schäfer: Simulation von Wartenetzen mit getakteter Übergabe, Diplomarbeit Nr. 33, FG Nachrichtentechnik, Universität Siegen (1982).

R. Schassberger, H. Daduna: The Time for a Round Trip in a Cycle of Exponential Queues, J. Assoc. Comput. Mach. 30 (1983), S. 146-150.

B. Schmidt: GPSS-FORTRAN, John Wiley & Sons, New York (1980).

W. Schmitt: Approximate Analysis of Markovian Queueing Networks with Priorities, 10th Intern. Teletraffic Congress, Montreal (1983), Congressbook, paper 1.3.3.

P. Schnupp: Rechnernetze: Entwurf und Realisierung, Walter de Gruyter, Berlin/New York (1978).

A. Schwanke: Wirksamkeit von Datenflußsteuerungs-Mechanismen in Paketvermittlungsnetzen, Studienarbeit Nr. 23, FG Nachrichtentechnik, Universität Siegen (1981).

M. Schwartz: Computer Communication Network Design and Analysis, Prentice-Hall, Englewood Cliffs, N.J. (1977).

M. Schwartz: Performance Analysis of the SNA Virtual Route Pacing Control, IEEE Trans. Commun. COM-30 (1982), S. 172-184.

M. Schwartz, S. Saad: Analysis of Congestion Control Techniques in Computer Communication Networks, Intern. Symposium on Flow Control in Computer Networks, Versailles (1979), S. 113-130, North-Holland Publishing Company (Ed. J. L. Grangé, M. Gien), Amsterdam/New York/Oxford.

M. Schwarz, T. E. Stern: Routing Techniques used in Computer Communication Networks, IEEE Trans. Commun. COM-28 (1980), S. 539-552.

P. J. Schweitzer, S. S. Lam: Buffer Overflow in a Store-and-Forward Network Node, IBM J. Res. Develop. 20 (1976), S. 542-550.

O. G. Soto, L. M. Miguez, I. G. Niemegeers: Modeling and Performance Evaluation of a Packet Switching Module in an X.25 Network, 10th Intern. Teletraffic Congress, Montreal (1983), Congressbook, paper 3.3.8.

D. E. Sproule, F. Mellor: Routing, Flow, and Congestion Control in the Datapac Network, IEEE Trans. Commun. COM-29 (1981), S. 386-391.

E. Stiefel: Einführung in die numerische Mathematik, 5. Auflage, Teubner Verlagsgesellschaft, Stuttgart (1976).

C. A. Sunshine: Efficiency of Interprocess Communication Protocols for Computer Networks, IEEE Trans. Commun. COM-25 (1977), S. 287-293.

J. Swoboda: Codierung zur Fehlerkorrektur und Fehlererkennung, R. Oldenbourg Verlag, München (1973).

R. Syski: Markovian Queues, Symposium on Congestion Theory, University of North Carolina (1964), S. 170-227, The University of North Carolina Press (Ed. W. L. Smith, W. E. Wilkinson), Chapel Hill.

L. Takács: Introduction to the Theory of Queues, Oxford University Press, New York (1962).

Y. Takahashi, H. Miyahara, T. Hasegawa: An Approximation Method for Open Restricted Queueing Networks, Operations Research 28 (1980), S. 594-602.

Y. Takahashi, N. Shigeta, T. Hasegawa: An Approximation Analysis for Congestion Control Scheme in Distributed Processing Systems, Intern. Conf. on Performance of Data Communication Systems and their Applications, Paris (1981), North-Holland Publishing Company (Ed. G. Pujolle), Amsterdam/New York/Oxford (1981), S. 345-354.

A. S. Tanenbaum: Computer Networks, Prentice-Hall, Englewood Cliffs, N. J. (1981).

K. Tanno, H. Kuniyoshi, T. Nakamura, R. Sato: A Flow Control Analysis Based on Measure of Power in Packet Switching Networks, National Telecommunication Conference, New Orleans (1981), S. E5.7.1/6.

H. C. Tijms, M. H. van Hoorn: Algorithms for the State Probabilities and Waiting Times in Single Server Queueing Systems with Random and Quasirandom Input and Phase-Type Service Times, OR Spektrum 2 (1981), S. 145-152.

P. Tran-Gia: Überlastprobleme in rechnergesteuerten Fernsprechvermittlungssystemen - Modellbildung und Analyse, Dissertation, Universität Siegen (1982).

P. Tran-Gia: Simulation of Instationary Processes for Performance Evaluations of Switching Systems, 1st European Simulation Congress, Aachen (1983), S. 362-367, Informatik-Fachberichte Band 71, Springer Verlag, Berlin/Heidelberg/New York/Tokyo.

R. A. Upton, S. K. Tripathi: An Approximate Transient Analysis of the M(t)/M/1 Queue, Performance Evaluation 2 (1982), S. 118-132.

P. Vogt: Auswirkung von hierarchischen Protokollebenen auf Durchsatz und Verzögerung in Packet Switching Datennetzen, Studienarbeit Nr. 37, FG Nachrichtentechnik, Universität Siegen (1983).

J. Walrand, P. Varaiya: Sojourn Times and the Overtaking Condition in Jacksonian Networks, Adv. Appl. Prob. 12 (1980), S. 1000-1018.

J. A. White, J. W. Schmidt, G. K. Bennett: Analysis of Queueing Systems, Academic Press, New York/San Francisco/London (1975).

W. Whitt: The Queueing Network Analyzer, Bell System Technical J. 62 (1983), S. 2779-2815.

W. Whitt: Performance of the Queueing Network Analyzer, Bell System Technical J. 62 (1983), S. 2817-2843.

G. Willmann: Lösungsalgorithmen für Warteschlangennetze mit Produktlösungsform, Diplomarbeit Nr. 38, FG Nachrichtentechnik, Universität Siegen (1983).

J. W. Wong: Distribution of End-to-End Delay in Message-Switched Networks, Computer Networks 2 (1978), S. 44-49.

J. W. Wong, S. S. Lam: Queuing Network Models of Packet Switching Networks, Part 1: Open Networks, Performance Evaluation 2 (1982), S. 9-21.

J. W. Wong, M. S. Unsoy: Analysis of Flow Control in Switched Data Networks, IFIP Conf. on Information Processing, North-Holland Publishing Company, Amsterdam (1977), S. 315-320.

T. P. Yum: The Design and Analysis of a Semidynamic Deterministic Routing Rule, IEEE Trans. Commun. COM-29 (1981), S. 498-504.

T. P. Yum, M. Schwartz: The Join-Biased-Queue Rule and its Application to Routing in Computer Communcation Networks, IEEE Trans. Commun. COM-29 (1981), S. 505-511.

CCITT-Empfehlungen

CCITT-Yellow Book, Part VIII.2/3, International Telecommunication Union, Geneva (1980).

CCITT-Empfehlungen der V-Serie und der X-Serie, Band 1: Datenpaketvermittlung - Internationale Standards, 4. Auflage (1981), Band 3: Datenübermittlungsnetze, 4. Auflage (1982), R. v. Decker's Verlag, G. Schenck, Heidelberg/Hamburg.

X.3: CCITT -Empfehlung X.3, Paketierungs-/Depaketierungs-Einrichtung (PAD - packet assembly/disassembly facility) in einem öffentlichen Datennetz.

X.21: CCITT-Empfehlung X.21, Schnittstelle zwischen Datenendeinrichtungen (DEE) und Datenübertragungseinrichtung (DÜE) für Synchronverfahren in öffentlichen Datennetzen.

X.25: CCITT-Empfehlung X.25, Schnittstelle zwischen Datenendeinrichtung (DEE) für Endeinrichtungen, die im Paketmodus in öffentlichen Datennetzen arbeiten.

X.28: CCITT-Empfehlung X.28, Schnittstelle zwischen DEE/DÜE für eine Start/Stop-Datenendeinrichtung, die eine Paketierung-/Depaketierungs-Einrichtung (PAD) eines öffentlichen Datennetzes im selben Land erreicht.

X.29: CCITT-Empfehlung X.29, Verfahren für den Austausch von Steuerinformationen und von Benutzerdaten zwischen einer Paket-DEE und einer Paketierungs-/Depaketierungs-Einrichtung (PAD).

X.75: CCITT-Empfehlung X.75, Steuerungsverfahren auf internationalen Verbindungsleitungen zwischen Datennetzen mit Paketvermittlung über Endverbindungen und Transitverbindungen für die Datenübermittlung.