KERNFORSCHUNGSZENTRUM

KARLSRUHE

April 1974 KFK 1972

Institut für Datenverarbeitung in der Technik

Modelle zur Ermittlung von Nachrichtenverzögerungen in Rechnernetzen

R. Senger

GESELLSCHAFT
FÜR
KERNFORSCHUNG M.B.H.

KARLSRUHE

GESELLSCHAFT FÜR KERNFORSCHUNG M. B. H.
KARLSRUHE

KERNFORSCHUNGSZENTRUM KARLSRUHE

KFK 1972

Institut für Datenverarbeitung
in der Technik

Modelle zur Ermittlung von Nachrichtenverzögerungen in Rechnernetzen

R. Senger

Gesellschaft für Kernforschung m.b.H, Karlsruhe

Kurzfassung:

Die Bedienungsqualität von Rechnernetzen wird entscheidend von der Leistungsfähigkeit des Rechnernetz-Kommunikationssystems beeinflußt, da die Verzögerungen im Kommunikationssystem sich unmittelbar auf die Antwortzeiten des Systems auswirken. Die vorliegende Arbeit versucht zu zeigen, inwieweit sich die bekannten Methoden der Warteschlangentheorie bei der Analyse und Synthese von Rechnernetz-Kommunikationssystemen anwenden lassen. Dazu wird insbesondere abgegrenzt, unter welchen Bedingungen Serienschaltungen und Netzwerke von Warteschlangen einer analytischen Behandlung zugänglich sind. Außerdem wird die Möglichkeit einer verfeinerten Modellbildung anhand eines Simulationsmodells für Netzwerke von Warteschlangen vorgestellt, mit dessen Hilfe mathematisch nicht erfaßbare Probleme behandelt werden können.

Abstract

Models for Estimation of Message Delays in Computer Networks

Computer network performance measured in terms of delays and responsetimes depends strongly on the performance of the communication system available for inter-computer communications. This paper reviews the applicability of Queuing Theory to analysis and synthesis of computer network communication systems. Conditions under which series and networks of queues are analytically tractable are discussed. As a way of more refined modelling, a simulation model for queuing networks was built, which proved to be helpful in analyzing problems where analytic solutions are not available.

Inhalt:

Wichtige Bezeichnungen

EW(X)	Erwartungswert der Zufallsgröße X
K_{ij}	Übertragungskanal zwischen den Vermittlungseinrichtungen V_i und V_j
k_{ij}	Kapazität des Kanals K_{ij}
L	Nachrichtenlänge
N_i	Nachricht
S	Bedienungszeit
s^2	Varianz (Schätzwert)
V_i	Vermittlungseinrichtung
Var(X)	Varianz der Zufallsvariablen X
VZ	Verzögerung (VZ=WZ+S)
WZ	Wartezeit
$\bar{x}$	Mittelwert einer Stichprobe (Schätzwert)
λ	Ankunftsrate
μ	Bedienungsrate
ρ	Verkehrsintensität
σ_s^2	Varianz der Bedienungszeit S
σ_z^2	Varianz der Zwischenankunftszeit Z

1. Einführung

Ein Rechnernetz besteht aus wenigstens zwei autonomen Rechensystemen, die selbständig Aufträge verarbeiten, und einem Kommunikationssystem. Letzteres dient zur Übermittlung von Nachrichten zwischen den Rechensystemen und besteht aus Vermittlungseinrichtungen und Kanälen (s. Bild 2). Hinsichtlich der Auslegung des Kommunikationssystems gibt es viele Variationsmöglichkeiten, sowohl was physikalische Einrichtungen als auch Organisationsprinzipien betrifft. Dazu sei jedoch auf [7] und [8] verwiesen. Hier soll nur kurz auf zwei alternative Nachrichtenübermittlungsverfahren, nämlich das sogenannte Line-Switching und das Message-Switching, eingegangen werden. Beim Line-Switching wird vor Beginn der Nachrichtenübermittlung eine Leitung über die Vermittlungseinrichtungen durchgeschaltet und während des ganzen Übertragungsvorgangs aufrechterhalten (vgl. z.B. Telefonverkehr). Beim Message-Switching dagegen wird eine Nachricht schon dann abgeschickt, wenn der Kanal zur nächsten Vermittlungseinrichtung frei wird. Dabei enthalten die Nachrichten Informationen über ihre Vorgeschichte und ihren Zielort. In den Vermittlungseinrichtungen werden sie in Warteschlangen gespeichert, und nach Prüfung und Quittierung der erhaltenen Nachricht wird unter Berücksichtigung des Netzzustands über die weitere Behandlung entschieden. Der Aufbau des Übertragungsweges erfolgt also erst während der Übertragung, und zwar schrittweise zwischen je zwei Vermittlungseinrichtungen. Auf diese Weise kann man eine Anpassung an aktuelle Störzustände und Lastverteilungen im Netz erreichen.

Wenn man sich nur für die Verzögerungen der Nachrichten bei der Übermittlung durch das Kommunikationssystem interessiert, kann man auf die Unterscheidung zwischen den beiden Übermittlungsverfahren verzichten, da jedem Line-Switching gewissermaßen ein Message-Switching zur Suche eines freien Übertragungswegs vorausgeht, bei dem sich wiederum Warteschlangen in den Vermittlungseinrichtungen bilden können. Die für die Analyse eines solchen Systems in Frage kommenden Hilfsmittel

der Warteschlangentheorie sind also in jedem Fall dieselben. Wenn im folgenden von Nachrichtenverzögerung die Rede ist, ist stets nur die Verzögerung durch den Aufenthalt in Warteschlangen und die Bearbeitung in Vermittlungseinrichtungen gemeint, nicht aber die Verzögerung aufgrund der endlichen Fortpflanzungsgeschwindigkeit der elektrischen Signale, deren Einfluß als vernachlässigbar betrachtet wird. Ferner wird zwischen der Wartezeit WZ, das ist die Zeit, die eine Nachricht in einer Warteschlange verbringt und der Verzögerung VZ, das ist die Verweilzeit einer Nachricht im Wartesystem (Wartezeit + Bedienung), unterschieden (vgl. auch Kap.7).

Als Modelle für die Untersuchung von Nachrichtenflüssen in Rechnernetz-Kommunikationssystemen bieten sich Warteschlangenmodelle an. Eine zentrale Rolle werden dabei erwartungsgemäß Serienschaltungen und Netzwerke von Warteschlangen spielen, deren Aufbau und Zusammenhang die Struktur des zugrundeliegenden Kommunikationssystems widerspiegeln. Aus diesem Grund werden die Möglichkeiten der Analyse von solchen komplexen Bedienungssystemen mit Hilfe der Warteschlangentheorie in Kap.3 ausführlich diskutiert.

2. Ein Kommunikationssystem-Modell

2.1. Modelldefinition

Im folgenden werden die Komponenten eines Kommunikationssystem-Modells beschrieben, das die Fähigkeit zum Message-Switching besitzen soll:

Nachrichten sind Informationsblöcke variabler Länge L [bit]. Sie enthalten u.a. Angaben über

- Ursprungsort
- Zielort
- Länge der Nachricht

Vermittlungseinrichtungen

besitzen die Fähigkeit, die Nachrichten durch Analyse der in diesen enthaltenen Steuerinformation durch das Netz zu ihrem Zielort zu lenken.

Die Zeit, die eine Nachricht N_i in einer Vermittlungseinrichtung verbringt, setzt sich zusammen aus der Wartezeit auf Vermittlung, der Bedienungszeit (Vermittlungszeit) und der Wartezeit, die die Nachricht in einer Kanalwarteschlange verbringt (s. Bild 3). Der Warteraum in allen Warteschlangen ist unbegrenzt und die Auswahl der nächsten zu bearbeitenden Nachricht erfolgt nach dem Prinzip "wer zuerst kommt, wird zuerst bedient".

Es wird angenommen, daß die Bedienungszeit in der Vermittlungseinrichtung für jede Nachricht eine negativ exponentiell verteilte Zufallsgröße ist:

$$P(S \leq t) = 1 - e^{-\mu t} \qquad t > 0 \qquad (1)$$

μ: Bedienungsrate

Für die Weiterleitung der Nachrichten soll folgendes gelten:

Das Netz bestehe aus n Vermittlungseinrichtungen (Knoten). Nach Bearbeitung durch Knoten V_i wird eine Nachricht mit der Wahrscheinlichkeit Θ_{ij} in die Warteschlange des Kanals zum Knoten V_j eingeordnet, oder sie verläßt das Kommunikationssystem mit der Wahrscheinlichkeit

$$1 - \sum_{j=1}^{n} \Theta_{ij}.$$

Kanäle

Die Kanäle übertragen die Nachrichten zwischen den Vermittlungseinrichtungen. Sie sind charakterisiert durch die Kanalkapazität k_{pq} [bit/Zeiteinheit]. Die Bedienungszeit eines Kanals K_{pq} für eine Nachricht der Länge l_i ist gegeben durch:

$$s_{K_{pq}}(N_i) = \frac{l_i}{k_{pq}} \qquad (2)$$

Nachrichtenströme

Es wird angenommen, daß an jeder Vermittlungseinrichtung sowohl neue Nachrichten in das Netz gelangen (von autonomen Rechensystemen) als auch dieses verlassen können. Die Ankunftsströme sollen stets Poisson-Ströme sein, d.h. die Wahrscheinlichkeit des Auftretens von k Nachrichten im Zeitintervall t ist gegeben durch:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad \lambda > 0 \qquad (3)$$

2.2. Bestimmung von Nachrichtenverzögerungen

Mit Hilfe der in 2.1. beschriebenen Kommunikationssystem-Komponenten (Kanäle und Vermittlungseinrichtungen) lassen sich beliebige, analog zu realen Rechnernetz-Kommunikationssystemen strukturierte Warteschlangen-Modelle aufbauen. Als Weg zur Bestimmung der Verzögerung beim Übermitteln einer Nachricht von V_i nach V_j über mehrere weitere Vermittlungseinrichtungen bietet sich unmittelbar die Aufteilung des Gesamtsystems in einfachere Bedienungssysteme, etwa vom Typ M/M/n, an. (Begriffe der Warteschlangentheorie s. Kap. 7). Wenn es gelingt, solche Teilsysteme festzulegen, die sich genauso verhalten, als ob sie isolierte Systeme, z.B. des obengenannten Typs wären, dann kann man die bekannten Ergebnisse der Warteschlangentheorie einfach übertragen. Der Erwartungswert EW (VZ_{ij}) der Verzöge-

rung einer Nachricht bei Übertragung von V_i nach V_j ergibt sich dann als Summe der Erwartungswerte der Verzögerungen an sämtlichen durchlaufenen Kanälen und Vermittlungseinrichtungen, EW (VZK_{pq}) und EW (VZV_l):

$$EW(VZ_{ij}) = \sum_{\{p,q\}\varepsilon W_1} EW(VZK_{pq}) + \sum_{l\varepsilon W_2} EW(VZV_l) \qquad (4)$$

W_1, W_2: Indexmengen, die den Weg von V_i nach V_j repräsentieren

Um die $EW(VZK_{pq})$ und $EW(VZV_l)$ bestimmen zu können, muß jedoch zunächst sichergestellt werden, daß sich die Teilsysteme wie unabhängige Systeme etwa des obengenannten Typs verhalten.

Die Unabhängigkeit der Teilsysteme wird jedoch offensichtlich dadurch gestört, daß die Bedienungszeiten an aufeinanderfolgenden Teilsystemen von einer Eigenschaft der Nachrichten, in diesem Fall der Nachrichtenlänge, abhängen. Insbesondere gilt dies natürlich dann, wenn die Länge der Nachrichten auf dem Weg durch das Kommunikationsnetz konstant ist. Kleinrock [14] hat gezeigt, daß eine analytische Lösung in diesem Fall an der Komplexität der auftretenden mathematischen Probleme scheitert, und dieses Problem durch folgenden Kunstgriff umgangen:

> Die Länge der Nachrichten sei während des Transports durch das Kommunikationsnetz nicht konstant, sondern werde jeweils beim Verlassen einer Vermittlungseinrichtung aus einer Exponentialverteilung neu bestimmt:
>
> $$P(L \leq l) = 1 - e^{-\mu l} \qquad (5)$$
>
> $1/\mu$: mittlere Nachrichtenlänge

Ein Kanal verhält sich dann wie ein M/M/1-Bedienungssystem, an dem der Erwartungswert der Verzögerung bekanntlich gege-

ben ist durch

$$EW(VZ) = \frac{1}{\mu-\lambda} \qquad (6)$$

μ: Bedienungsrate
λ: Ankunftsrate

Zusammen mit (2):

$$s_{k_{pq}}(N_i) = \frac{l_i}{k_{pq}}$$

ergibt sich daraus:

$$EW(VZK_{ij}) = \frac{1}{\mu k_{ij}-\lambda} \qquad (7)$$

k_{ij}: Kapazität des Kanals von V_i nach V_j

Für die Verzögerung in den Vermittlungseinrichtungen, $EW(VZV_i)$, läßt sich nach Voraussetzung (1) die Formel für das M/M/1-System (6) direkt anwenden.

Damit sind die Terme aus (4) ermittelt und jedes aus den in 2.1. vorgestellten Komponenten zusammengesetzte Kommunikationssystem-Modell kann durch getrennte Behandlung der Teilsysteme (Vermittlungseinrichtungen bzw. Kanäle) analysiert werden. Zu beachten ist jedoch, daß auf diesem Weg nur der Erwartungswert der Verzögerung bei Übertragung über mehrere Knoten, nicht aber höhere Momente der Verteilung von VZ_{ij} bestimmt werden können. Dazu wäre die Unabhängigkeit aller Verzögerungen VZK_{pq} und VZV_l an allen Teilsystemen, die zu unterscheiden ist von der obengenannten "unabhängigen Arbeitsweise" der Teilsysteme, nachzuweisen. Auf diese Probleme wird in Kap.3 näher eingegangen.

2.3. Diskussion der Gültigkeit des Modells

Das vorgestellte Modell eines Rechnernetz-Kommunikationssystems besitzt einen hohen Abstraktionsgrad. Es stellt nicht etwa ein Abbild eines bestehenden oder geplanten Systems dar, sondern es soll vor allem dazu dienen, die Möglichkeiten und Grenzen der Warteschlangentheorie bei der Analyse von Rechnernetz-Kommunikationssystemen auf einer möglichst allgemeinen Basis zu erläutern.

Trotzdem darf die Frage nach der Gültigkeit des Modells bzw. nach seiner Aussagekraft im Hinblick auf reale Systeme nicht unbeachtet bleiben.

Zunächst ist die Annahme von Poissonströmen als Eingangsströme an den Knoten des Netzes zu rechtfertigen.

Man kann zeigen [11], daß die Summe einer großen Zahl von Auftragsströmen, von denen jeder auf die Summe nur einen geringen Einfluß hat, unabhängig von der Art der Teilströme dem Poissonstrom ähnlich ist. Auf Rechnernetzkommunikationssysteme übertragen bedeutet dies:

Resultieren die Nachrichten aus Benutzeraufträgen, die an verschiedenen Stellen unabhängig voneinander in das Rechnernetz geschleust werden, so wird der Poissonstrom eine akzeptable Näherung für den realen Strom darstellen. Diese Situation ist in der Realität häufig anzutreffen und die daraus resultierende Poisson-Charakteristik kann für viele Fälle durch entsprechendes statistisches Datenmaterial belegt werden.

Weitaus kritischer ist die Voraussetzung der negativ exponentiell verteilten Bedienungszeiten, die oft, wie auch in unserem Fall (Kanäle), nicht deshalb gewählt wird, weil dadurch eine realitätsnahe Modellierung gelingt, sondern weil mit dieser Voraussetzung die mathematische Behandlung von Bedienungssystemen vereinfacht bzw. überhaupt erst ermöglicht wird. Um dies zu verdeutlichen, sei eine charakteristische

Eigenschaft eines Bedienungssystems mit exponentiell verteilter Bedienungszeit, die "Vergeßlichkeit", hier kurz erläutert.

Exponentiell verteilte Bedienungszeit bedeutet, daß das Auftreten von Ereignissen im Intervall $(t, t+\Delta t)$ nur vom Zustand des Systems zum Zeitpunkt t, nicht aber von der Vorgeschichte, die zu diesem Systemzustand geführt hat (Intervall $(o,t]$), abhängt. Das heißt z.B., daß die restliche Bedienungszeit eines Auftrags zu jedem Zeitpunkt unabhängig von der bereits verstrichenen Bedienungszeit ist:

Ist die Verteilung der Bedienungszeit gegeben durch

$$P(S \leq s) = F(s) = 1-e^{-\mu s}$$

und ist

$$p_a(t) = P \{\text{Bedienung, die schon die Zeit a andauert, währt mind. noch die Zeit t}\}$$

so gilt (vgl. [11]):

$$p_o(t) = e^{-\mu t}$$

$$p_o(a) = e^{-\mu a}$$

$$p_o(a+t) = e^{-\mu(a+t)}$$

da stets $p_o(a+t) = p_o(a) \cdot p_a(t)$

folgt $e^{-\mu(a+t)} = e^{-\mu a} \cdot p_a(t)$

und damit $p_a(t) = e^{-\mu t} = p_o(t)$

Besonders unrealistisch scheint die Annahme, daß die Länge einer Nachricht beim Durchlaufen des Netzes nicht konstant bleibt (vgl. 2.2.). Andererseits hat Kleinrock [14] mit Hilfe der Simulation gezeigt, daß man auch mit dieser Annahme bei zunehmender Vermaschung des Netzes brauchbare Ergebnisse erzielen kann.

3. Warteschlangentheoretischer Hintergrund

In diesem Kapitel wird auf die Problematik der Zerlegung eines Kommunikationssystems der in Bild 2 gezeigten Art in isolierte, analysierbare Teilsysteme eingegangen. Bild 4 zeigt das Grundmodell einer Serienschaltung von zwei Bedienungssystemen A und B, an dem zunächst die Voraussetzungen für eine Zerlegung erläutert werden sollen. Da der Ausgangsstrom S_A von System A identisch mit dem Eingangsstrom S_B für System B ist, wird sich der Versuch, System (A,B) zu analysieren, zunächst auf die Frage nach der Art des Ausgangsstroms S_A von System A konzentrieren. Von besonderer Bedeutung ist dabei die Frage, ob S_A aufgrund seiner statistischen Eigenschaften wieder als Teilkomponente eines mathematischen Modells, nämlich des Systems B, dienen kann.

Dies ist keineswegs selbstverständlich, denn:

> Die grundlegende Voraussetzung für die Analyse nahezu aller Warteschlangenmodelle ist die Rekurrenz des Eingangsstroms, d.h. die Zwischenankunftszeiten $z_i = t_i - t_{i-1}$ zwischen dem Eintreffen zweier unmittelbar aufeinanderfolgender Aufträge zu den Zeitpunkten t_{i-1} bzw. t_i müssen unabhängig identisch verteilt sein. (Ausnahmen: siehe z.B. [10]).

Das bedeutet: Eine Mindestvoraussetzung für die Zerlegung in Teilsysteme ist die Rekurrenz des Ausgangsstroms S_A bzw. bei komplexeren Systemen die Rekurrenz aller Ausgangsströme, die gleichzeitig Eingangsströme für weitere Teilsysteme sind.

3.1 Der Ausgangsstrom einer Bedienungsstation

Um die Frage nach der Art des Ausgangsstroms zu beantworten, seien hier die diesbezüglichen Ergebnisse der Warteschlangentheorie zusammengefaßt.

Burke [1] hat gezeigt, daß für die Bedienungsstation vom Typ M/M/n gilt:

> Der Ausgangsstrom der Bedienungsstation M/M/n ist im stationären Fall ein Poissonstrom, dessen Parameter mit dem des Eingangsstroms übereinstimmt.

Wählt man jedoch eine andere Verteilung für den Ankunftsstrom und die Bedienungszeit (z.B. eine χ^2-Verteilung mit vier Freiheitsgraden, was eine nur leichte Veränderung bedeutet [17], so ist der Ausgangsstrom i.a. nicht mehr mit dem Eingangsstrom statistisch identisch. Finch [9] zeigte sogar, daß die exponentiell verteilte Bedienungszeit und unbeschränkter Warteraum notwendige Bedingungen für die Rekurrenz des Ausgangsstroms eines M/M/2-Systems sind.
Man kann also zusammenfassend feststellen, daß der Ausgansstrom einer Bedienungsstation i.a. nur dann rekurrent ist, wenn die Bedienungszeit exponentialverteilt ist, wenn es auch für einige spezielle Bedienungsstationstypen mit besonderen Voraussetzungen, wie z.B. M/G/∞, gelungen ist, zu zeigen, daß ihr Ausgangsstrom ebenfalls ein Poissonstrom ist [16].

3.2 Überlagerung und Aufspaltung von Poissonströmen [5]

Eine Überlagerung zweier Poissonströme P_1 und P_2 mit der Ankunftsrate λ_1 bzw. λ_2 führt zu einem resultierenden Poissonstrom P_3 mit Ankunftsrate $\lambda_3 = \lambda_1+\lambda_2$. Dasselbe gilt entsprechend für eine größere Zahl von sich überlagernden Poissonströmen (Bild 5).

Auch bei der Aufspaltung eines Poissonstroms in zwei oder mehrere Teilströme sind diese unter einer gewissen Voraussetzung wieder Poissonströme (Bild 5). Voraussetzung ist, daß der weitere Weg eines Auftrages am Verzweigungspunkt zufällig mit der Wahrscheinlichkeit r bzw. 1-r ausgewählt wird, wenn zwei alternative Wege vorliegen. Ist die An-

kunftsrate des ungeteilten Stroms P_1 gleich λ_1, so besitzen die Teilströme P_2 und P_3 die Ankunftsrate $\lambda_2=r\lambda_1$ bzw. $\lambda_3=(1-r)\lambda_1$.

Entscheidend dafür, daß wieder Poissonströme entstehen, ist, daß die Aufteilung der Aufträge auf die verschiedenen möglichen Wege durch zufällige und unabhängige Auswahl erfolgt:

Wählt man z.B. für jeden zweiten Auftrag denselben Weg, so resultiert daraus kein Poissonstrom, sondern ein Erlang-2-Strom.

3.3 Netzwerke von Bedienungsstationen

J.R. Jackson gibt in 12 an, unter welchen Bedingungen man die Teilsysteme eines Netzwerkes von Bedienungssystemen wie unabhängige Systeme behandeln kann. Seine Ergebnisse seien an dieser Stelle zusammengefaßt.

Wir betrachten eine Anzahl von Bedienungsstationen S_1, S_2, ...S_M (M>2) (s. Bild 6), für die folgende Voraussetzungen erfüllt sind:

- Aufträge können sowohl von außerhalb des Gesamtsystems als auch von anderen Teilsystemen zum System S_m gelangen.
- Nach der Beendigung ihrer Bedienung im System S_m gehen die Aufträge (sofort) zum System S_k zur weiteren Bearbeitung mit der Wahrscheinlichkeit Θ_{km}, oder verlassen das Gesamtsystem mit der Wahrscheinlichkeit $1-\sum_k \Theta_{km}$.

Weiter bedeuten:

M	Zahl der Teilsysteme
s_m	Zahl der (parallelen) Bedienungsschalter im System S_m
λ_m	Ankunftsrate des Auftragsstroms, der von außerhalb des Gesamtsystems in das Teilsystem S_m mündet.

μ_m Bedienungsraten der Bedienungsschalter im System S_m

Γ_m Ankunftsrate des Gesamtankunftsstroms am System S_m von inner- oder außerhalb des Netzes.

$p(k_1,k_2,\ldots,k_m) = P\{(k_1$ Aufträge in $S_1)$ $(k_2$ Aufträge in $S_2)$ ... $(k_m$ Aufträge in $S_m)\}$

$p_k^i = P\{k$ Aufträge im Teilsystem $S_i\}$

Alle von außerhalb kommenden Ströme seien Poissonströme und alle Bedienungszeiten exponentiell verteilt. Mit den oben getroffenen Vereinbarungen ergibt sich

$$\Gamma_m = \lambda_m + \sum_k \Theta_{mk} \cdot \Gamma_k$$

Für ein Bedienungsnetz mit den genannten Voraussetzungen gilt im stationären Fall ($\Gamma_m < \mu_m, s_m$ für $m=1,2,\ldots,M$):

$$p(k_1,k_2,\ldots,k_M) = p_{k_1}^1 \; p_{k_2}^2 \; \ldots \; p_{k_M}^M$$

wobei die p_k^m gegeben sind durch (vgl. [19]):

$$p_k^m = \begin{cases} p_0^m \dfrac{\left(\dfrac{\Gamma_m}{\mu_m}\right)^k}{k!} & \text{für } k=0,1,\ldots,s_m \\ p_0^m \dfrac{\left(\dfrac{\Gamma_m}{\mu_m}\right)^k}{n_m!\,(n_m)^{k-s_m}} & \text{für } k=s_m,s_{m+1}\ldots \end{cases}$$

(mit $m=1,2,\ldots,M$; $k=0,1,2,\ldots$)

Damit ist eine Verallgemeinerung der Aussagen von 3.1 gelungen:

In einem Netz aus Bedienungsstationen des Typs M/M/n, wie es hier beschrieben wurde, verhalten sich die Teilsysteme so, als ob sie unabhängige, isolierte Teilsysteme des Typs M/M/n wären. Entscheidend dafür ist, daß alle Bedienungszeiten exponentialverteilt sind und daß der weitere Weg eines Auftrags nach der Bedienung an einem Teilsystem zufällig bestimmt wird. Aus dieser Sicht werden auch die Gründe für die speziellen Voraussetzungen bei der Definition des mathematischen Modells in 2.1 deutlich. Dieses Modell ist so ausgelegt, daß eine Zerlegung in Teilsysteme möglich ist. Für jedes Teilsystem kann der Erwartungswert der Verzögerung bestimmt und für einen Übertragungsweg über mehrere Kanäle und Vermittlungseinrichtungen können die Erwartungswerte zum Erwartungswert der Gesamtverzögerung aufsummiert werden. Höhere Momente der Verteilung der Gesamtwartezeit und der Gesamtverzögerung können jedoch i.a. nich berechnet werden, da die Wartezeiten und Verzögerungen eines Auftrags an den Teilsystemen einer Übertragungsstrecke nicht unabhängig sind. (Ausnahme: Serienschaltung von M/M/1 -Systemen). |2,3,17,18|

4. Ansätze für eine verfeinerte analytische Modellbildung

Für einige Spezialfälle gelingt auch bei nicht exponentiell verteilten Bedienungszeiten eine mathematische Analyse. Zum Beispiel hat Friedmann [10] gezeigt, wie man Serienschaltungen von Wartesystemen des Typs G/D/n durch stufenweises Reduzieren untersuchen kann. Dabei ist der Eingangsstrom beliebig; er muß also insbesondere nicht rekurrent sein.

Außerdem kann ein Modell, das exponentialverteilte Bedienungszeiten voraussetzt, u.U. auch Schlüsse auf das Verhalten eines Systems mit nicht exponentialverteilten Bedienungszeiten zulassen. Die Exponentialverteilung ist ein Spezialfall der Erlangverteilung. Unter den Erlangverteilungen stellt sie als

Bedienungszeitverteilung den ungünstigsten Fall bezüglich der zu erwartenden Verzögerungen dar. Der Vorteil der Erlangverteilung besteht darin, daß sie eine weit bessere Anpassung an reale Bedienungszeitverteilungen gestattet als die Exponentialverteilung.

Die Dichte einer Erlangverteilung ist gegeben durch

$$f(x) = \frac{(\lambda K)^k x^{k-1}}{(k-1)!} e^{-\lambda k x} \qquad x \geq 0$$

mit $\lambda > 0$
und $K \in |N$

Im folgenden wird erläutert, wie man die exponentialverteilte Bedienungs- oder Zwischenankunftszeit als "schlechtesten Fall" unter den Erlangverteilungen zur Gewinnung von oberen Schranken für die Verzögerung von Aufträgen nutzen kann. Als Beispiel dient hier die Wartezeit, für die Verzögerung (Wartezeit + Bedienung) gilt jedoch entsprechendes.

Ein Maß für die Unregelmäßigkeit der Verteilung der Zufallsvariablen X ist der Variationskoeffizient

$$C^2 = \frac{VAR(X)}{(EW(X))^2}$$

Für eine Erlangverteilung mit Parameter k läßt er sich ausdrücken als

$$C^2 = \frac{1}{k}$$

und es gilt $0 \leq C^2 \leq 1$.

Die Exponentialverteilung (k=1) hat unter allen Erlangverteilungen den höchsten Variationskoeffizienten $C^2=1$.

Ein Vergleich der bekannten Formeln (vgl. Kap. 7) für die Erwartungswerte der Wartezeit im System M/M/1

$$EW(WZ)_M = \frac{\lambda}{\mu(\mu - \lambda)}$$

und im System $M/E_k/1$

$$EW_M > EW_k = \frac{(k+1)\ \lambda}{2k\mu(\mu-\lambda)}$$

zeigt, daß zunächst $E_M > E_k$ für $k = 2, 3, \ldots$

und sogar $EW_k > EW_{k+1}$

gilt, d.h., daß mit zunehmender Regelmäßigkeit die Erwartungswerte für die Wartezeit kleiner werden. Die Exponentialverteilung, die unregelmäßigste unter allen Erlangverteilungen, stellt den ungünstigsten Fall dar.

Entsprechend einfache Betrachtungen sind nicht möglich, wenn man zeigen will, daß die Exponentialverteilung auch als Verteilung der Zwischenankunftszeit zu einem ungünstigsten Fall für die Wartezeit führt. Der Grund dafür ist, daß Formeln ähnlich unkomplizierter Bauart wir bei den Systemen M/GI/1 nicht vorhanden sind. Erlang-verteilte Zwischenankunftszeiten ($k>1$) bedeuten jedoch, daß der Ankunftsstrom regelmäßiger wird ($C^2<<1$). Für den Extremfall konstanter Ankunftsintervalle läßt sich eine einfache Formel angeben. In [21] werden die Systeme M/M/1, M/D/1, D/M/1 und D/D/1 bezüglich des Erwartungswerts der Wartezeit miteinander verglichen. Dabei ergibt sich (für $0<\rho<1$):

$$EW(WZ)_{M/M/1} \geq EW(WZ)_{M/D/1} \geq EW(WZ)_{D/M/1} \geq EW(WZ)_{D/D/1}$$

Es zeigt sich außerdem, daß der Übergang vom System M/M/1 auf ein System M/D/1 nahezu die gleiche Auswirkung hat wie der Übergang auf ein System D/M/1:

$$EW(WZ)_{M/M/1} \geq EW(WZ)_{M/D/1} \approx EW(WZ)_{D/M/1} \geq EW(WZ)_{D/D/1}$$

Wenn eine Bedienungsstation voll ausgelastet ist, dann nimmt die Verteilung der Abgangsintervalle die Verteilung der Bedienungszeit an. Bei einer Serienschaltung von Systemen mit Erlang-verteilter Bedienungszeit kann man deshalb bei großer Verkehrsintensität ρ (<1) und bei großem Parameter k (der Erlang-Verteilung) damit rechnen, daß ein unregelmäßiger Auftragsstrom im System regelmäßiger wird und damit auch die Wartezeiten mit zunehmender Stufenzahl kürzer werden.

Eine weitere Hilfe bei der Suche nach Näherungslösungen bietet die Ungleichung von Kingman [13] für Wartesysteme vom Typ GI/G/n:

$$EW(WZ) \leq \frac{\sigma_s^2 + \sigma_z^2}{2(\frac{1}{\lambda} - \frac{1}{\mu})}$$

mit σ_s^2 : Varianz der Bedienunszeit

σ_z^2 : Varianz der Zwischenankunftszeit

Aus dieser Beziehung kann man sofort entnehmen, daß die obere Schranke für den Erwartungswert der Wartezeit mit größer werdener Regelmäßigkeit der Zwischenankunfts- oder Bedienungszeitverteilung bei konstanter Ankunftsrate kleiner wird.

5. Verfeinerte Modellbildung mit Hilfe der Simulation

Die Diskussion der Gültigkeit des in Kap. 2 vorgestellten Modells eines Rechnernetz-Kommunikationssystems hat gezeigt, daß man, um zu analytisch lösbaren Modellen zu kommen, unter Umständen Restriktionen in Kauf nehmen muß, die die geforderte Realitätsnähe des Modells in Frage stellen. Dies gilt in noch stärkerem Maße, wenn zu einem konkreten System mit festgelegten Eigenschaften ein Modell zu bilden ist.
Als Alternative zu mathematischen Modellen bietet sich die Modellbildung mit Hilfe der digitalen Simulation an [15]. Da bei dieser Methode die Ergebnisse nicht durch rechnerische Lösung von Modellgleichungen, sondern durch Experimente am Modell und anschließende Auswertung der anfallenden Daten gewonnen werden, entfällt die Notwendigkeit, das Modell so zu konstruieren, daß es einer analytischen Lösung zugänglich ist.

Zu dem in Kap. 2 vorgestellten Modell wurde ein entsprechendes Simulationsmodell entworfen und in der Sprache SIMULA implementiert [20], um

- Erfahrungen mit der Simulation von Wartesystemen zu sammeln
- Vergleichswerte für die mit Hilfe des mathematischen Modells erhaltenen Ergebnisse zur Verfügung zu stellen
- eine verfeinerte Modellbildung durch Umgehung der genannten Restiktionen (z.B. exponentiell verteilte Bedienungszeiten) zu erreichen.

5.1 Die Komponenten des Simulationsmodells

Die Ansatzpunkte für die gewünschte verfeinerte Modellbildung ergeben sich aus den in Kap. 2 beschriebenen Idealisierungen, deren wichtigste hier wiederholt seien:

1. Die Auswahl des weiteren Weges einer Nachricht erfolgt zufällig.

2. Die Bedienungszeiten in den Vermittlungseinrichtungen sind exponentialverteilt.

3. Die Länge jeder Nachricht wird beim Verlassen einer Vermittlungseinrichtung jedesmal neu aus einer Exponentialverteilung bestimmt.

Das Simulationsmodell besteht aus denselben Komponenten wie das in Kap. 2 beschriebene mathematische Modell, also Vermittlungseinrichtungen, Kanäle usw. Hinzu kommt ein neuer Typ von Teilsystemen, die die autonomen Rechensysteme repräsentieren und die Aufgabe haben, Nachrichten zu erzeugen und über das Kommunikationssystem auszutauschen. Bild 7 zeigt die Umgebung eines solchen autonomen Rechensystems.

Um eine verfeinerte Modellbildung zu erreichen, wurde das Kommunikationssystem wie folgt modifiziert:

1. Der Weg der Nachrichten durch das Kommunikationssystem wird nicht durch zufälliges Auswählen eines Kanals an den Vermittlungseinrichtungen bestimmt, sondern durch "echtes" Message-Switching aufgrund der in den Nachrichten (in Form eines Nachrichten-Kopfs) enthaltenen Steuerinformation anhand einer Routing-Table ermittelt. Die Routing-Table ist in jeder Vermittlungseinrichtung vorhanden. Aus ihr kann entnommen werden, welches der jeweils nächste Vermittlerknoten auf dem Weg zum Bestimmungsort eines Auftrags ist.

2. Die Bedienungszeiten in den Vermittlungseinrichtungen werden aus einer Erlang-Verteilung (s.Kap.4) bestimmt.

3. Die Länge der Nachrichten wird nicht mehr beim Verlassen jeder Vermittlungseinrichtung neu festgelegt, sondern während der gesamten Übertragung beibehalten. Die Bedienungszeit in den Kanälen

kann wahlweise abhängig von der Länge einer Nachricht ($s_{Kpq}=l_i/k_{pq}$, vgl. 2.1) oder unabhängig davon aus einer Erlangverteilung bestimmt werden.

Das Klassen-Konzept von SIMULA ermöglicht einen modularen Programmaufbau und unterstützt damit das bereits in Kap. 2 genannte Bestreben, das Modell so flexibel auszulegen, daß aus den Grundelementen Vermittlungseinrichtung, Kanal usw. beliebig strukturierte Kommunikationsnetze aufgebaut werden können. Das implementierte Simulationsprogramm erzeugt unter Berücksichtigung der Eingabedaten ein Netz aus den obengenannten Elementen mit gewünschten Eigenschaften. Eingabedaten sind u.a.:

- Zahl der autonomen Rechensysteme und Vermittlungseinrichtungen
- Verbindungsmatrix
- Nachfolgermatrix (Routing-Tables)
- Verkehrsmatrix (Intensität der Nachrichtenströme)

5.2 Gewinnung und Auswertung der Simulationsdaten

Es ist von Vorteil, wenn man bereits beim Entwurf der Simulationsexperimente Art und Umfang der geplanten statistischen Datenanalyse berücksichtigt. Dann ist zu entscheiden, welche Daten vom Simulationsmodell erzeugt werden sollen und wie diese für die anschließende statistische Analyse aufzubereiten sind, die die eigentliche Auswertung darstellt. Der für das in 5.1 beschriebene Modell eingeschlagene Weg soll hier kurz erläutert werden.

Die statistische Analyse der Simulationsdaten konzentrierte sich auf die Bestimmung von Mittelwerten von Systemparametern und deren Absicherung durch Konfidenzintervalle.

Das Ziel der durchgeführten Simulationsexperimente (s. 5.3) war die Bestimmung der Gesamtverzögerung der Nachrichten bei Übertragung zwischen Rechensystemen über eine Serie von Vermittlungseinrichtungen und Kanälen. Zu diesem Zweck ist bei der Erzeugung des Rechnernetz-Simulationsmodells durch Eingabe entsprechende Daten (s. 5.1) eine gewünschte Teststrecke festzulegen. Die Gesamtwartezeiten und Gesamtverzögerungen der Nachrichten, die diese Teststrecke passieren, werden akkumuliert und ausgegeben. Weiterhin werden die Mittelwerte der Wartezeiten und Verzögerungen an jedem Kanal und jeder Vermittlungseinrichtung ermittelt, um einen Überblick über das Verhalten des Kommunikationssystems zu gewinnen.

Neben der Aufgabe, die Einschwingperiode bzw. die Stationarität (siehe Kap. 7) des Systems festzustellen, auf die hier nicht näher eingegangen werden soll, tritt bei der Simulation von Warteschlangen vor allem das Problem auf, Vertrauensintervalle für Mittelwerte anzugeben, die i.a. aus autokorrelierten Stichprobenwerten berechnet wurden, wie folgende Überlegung zeigt:

Für eine einfache Bedienungsstation mit einem Schalter und natürlicher Warteordnung ("wer zuerst kommt, wird zuerst bedient") gilt:

$$wz_{i+1} = \max(wz_i + s_i - z_i, o)$$

wobei wz_i = Wartezeit des i-ten Auftrages

s_i = Bedienungszeit des i-ten Auftrages

z_i = Zeit zwischen der Ankunft des i-ten und des (i+1)-ten Auftrages

Wegen der deshalb zu erwartenden starken Korrelation der Wartezeiten aufeinanderfolgender Aufträge kann die bekannte Formel für die Schätzung der Varianz des Mittelwertes einer Stichprobe

$$Var(\overline{X}) = \frac{s^2}{n} \qquad (1)$$

mit s^2 : Varianz der Stichprobe

n : Stichprobenumfang

nicht unmittelbar angewandt werden, da dadurch die tatsächliche Varianz i.a. unterschätzt wird.

Eine einfache, aber auch sehr aufwendige Methode, um zu unabhängigen simulierten Daten zu kommen, besteht darin, jeden Simulationslauf mit verschiedenen Anfangswerten zu wiederholen, bis ein Stichprobenumfang erreicht ist, der es gestattet, den Mittelwert mit der gewünschten Genauigkeit zu bestimmen.

Eine andere Möglichkeit, von der hier Gebrauch gemacht wurde, besteht darin, für eine Stichprobe nicht mehrere, sondern nur einen (längeren) Simulationslauf durchzuführen und die anfallenden Daten so aufzubereiten, daß man nahezu unabhängige Stichprobenwerte erhält.

So werden z.B. in dem hier beschriebenen Simulationsmodell an den Teilsystemen nicht die Attribute aller Nachrichten erfaßt, sondern jeweils nur die jeder n-ten Nachricht ($n \geq 5$) um den Einfluß der Autokorrelation zu dämpfen.

Die so bestimmten Meßwerte werden als unabhängig betrachtet und die Vertrauensintervalle für Mittelwerte aufgrund von (1) bestimmt durch:

$$\bar{x}-\lambda_{Q\%} \cdot \frac{s}{\sqrt{n}} \leq m \leq \bar{x}+\lambda_{Q\%} \cdot \frac{s}{\sqrt{n}} \qquad (2)$$

mit $\bar{x}$: geschätzter Mittelwert
s : geschätzte Streuung
n : Stichprobenumfang
$\lambda_{Q\%}$: Q%-Grenze der (0,1)-Normalverteilung

Die Entscheidung, wie groß der Abstand zwischen den erfaßten Nachrichten gewählt wird, wird durch Korrelationsmessungen bei Pilotläufen ermöglicht. Den offensichtlichen Nachteil der genannten Methode, daß nämlich ein großer Teil der Daten nicht zur Auswertung verwendet wird, vermeidet ein anderes Verfahren [6], das zur Absicherung der Mittelwerte der Gesamtverzögerung bzw. Gesamtwartezeit auf der Teststrecke (s.o.) angewandt wird. Dabei werden die aus einem Simulationslauf gewonnenen Daten in Intervalle gleicher Länge eingeteilt, so daß man annehmen kann, daß die Korrelation zwischen nicht aufeinanderfolgenden Intervallen vernachlässigbar ist (was im Einzelfall wieder durch Korrelationsmessungen nachzuprüfen ist).

Die Einschwingperiode des Systems wird dadurch eliminiert, daß die während dieser Phase des Simulationslaufs anfallenden Daten nicht zur Auswertung herangezogen werden [6].

Mit Hilfe der Mittelwerte der Intervalle wird die Varianz des Gesamtmittelwerts abgeschätzt durch:

$$\mathrm{Var}(\bar{X}) < \frac{s^2}{10} \quad (1+ \frac{2r}{1-r}) \qquad (3)$$

wobei s^2 : geschätzte Varianz der Intervallmittelwerte

r : geschätzte Korrelation unmittelbar aufeinanderfolgender Intervalle

Die Bestimmung der Konfidenzintervalle erfolgt wieder nach (2).

Aufgrund der Konfidenzintervalle läßt sich angeben, welcher relative Fehler

$$f = \frac{|\bar{x} - m|}{\bar{x}}$$

mit einer Sicherheitswahrscheinlichkeit Q% höchstens auftritt.

5.3 Simulationsbeispiele

Mit dem in 5.1 beschriebenen Simulationssystem können, wie bereits erwähnt, beliebige Netzkonfigurationen mit gewünschten Eigenschaften erzeugt werden. An den folgenden Beispielen von mit dem System durchgeführten Simulationsexperimenten soll dies verdeutlicht werden. Außerdem wird erneut die in 2.3 diskutierte Frage der Validität des mathematischen Modells aufgegriffen. Zu diesem Zweck werden die Simulationsergebnisse den Ergebnissen der mathematischen Analyse gegenübergestellt, die im Falle der "Nichtlösbarkeit" des mathematischen Modells, also z.B. bei nicht exponentiell verteilten Bedienungszeiten, unter Berücksichtigung der Vorschläge von Kap. 4 ermittelt wurden.

In den Beispielen bedeuten:

λ	Ankunftsrate
μ	Bedienungsrate
ρ	Verkehrskoeffizient $\frac{\lambda}{\mu}$
EW(WZ)	Erwartungswert der Wartezeit, mathemat. Modell
EW(WZ)	Erwartungswert der Verzögerung, mathemat. Modell
$\bar{x}_{WZ}$	Mittelwert der Wartezeit, Simulationsmodell
$\bar{x}_{VZ}$	Mittelwert der Verzögerung, Simulationsmodell

Die Wahl der Systemparameter wurde so getroffen, daß unter Berücksichtigung der Stationaritätsbedingung $\rho<1$ eine gute Auslastung ($\rho\approx 0,8$) erzielt wurde und die Berechnung der Erwartungswerte (Warteschlangen-Formeln siehe Kap. 7) möglichst einfach war. Die Länge der Simulationsläufe war in allen Fällen so bemessen, daß für den relativen Fehler f (siehe 5.2) galt:

$f < 10\ \%$ mit einer Sicherheitswahrscheinlichkeit von 99 %.

Beispiel 1

Ein Poisson-Nachrichtenstrom mit $\lambda=10$ wird von einem Rechensystem R_1 über 6 Vermittlungseinrichtungen und 5 Kanäle zu einem Rechensystem R_2 übertragen (s. Bild 8). Die Bedienungszeit der Kanäle ist unabhängig exponentialverteilt mit $\mu=12$. Daraus ergibt sich der Verkehrskoeffizient $\rho=0,8\overline{3}$. In den Vermittlungseinrichtungen erfolgt keine Verzögerung.

Ergebnisse:

a) Wartezeiten an den Kanälen: EW(WZ)=0,4167 (math. Modell)

Kanal	$\bar{x}_{wz}$	Varianz	$EW(WZ)-\bar{x}_{wz}$	$\frac{EW(WZ)-\bar{x}_{wz}}{EW(WZ)}$ (%)
K_{12}	0,4073	0,2261	0,0093	2,2
K_{23}	0,4621	0,2551	-0,0455	-10,9
K_{34}	0,4561	0,3041	-0,0395	- 9,5
K_{45}	0,4456	0,2754	-0,0290	- 7,0
K_{56}	0,3720	0,1590	0,0460	10,7

b) Verzögerungszeiten an den Kanälen: EW(VZ)=0,5000 (math. Modell)

Kanal	$\bar{x}_{vz}$	Varianz	$EW(VZ)-\bar{x}_{v}$	$\frac{EW(VZ)-\bar{x}_{vz}}{EW(VZ)}$ (%)
K_{12}	0,4947	0,2313	0,0053	1,1
K_{23}	0,5471	0,2603	-0,0471	-9,4
K_{34}	0,5422	0,3094	-0,0422	-8,4
K_{45}	0,5294	0,2847	-0,0294	-5,9
K_{56}	0,4535	0,1772	0,0465	9,3

c) Gesamtwartezeiten und Gesamtverzögerungen auf der Teststrecke $R_1 \rightarrow R_2$

Wartezeiten:

$\bar{x}_{wz}$	EW(WZ)	$EW(WZ)-\bar{x}_{wz}$	$\frac{EW(WZ)-\bar{x}_{wz}}{EW(WZ)}$ (%)
2.150	2.083	-0,067	-3,2

Verzögerungen:

$\bar{x}_{wz}$	EW(VZ)	$EW(VZ)\ \bar{x}_{vz}$	$\frac{EW(VZ)-\bar{x}_{vz}}{EW(VZ)}$ (%)
2.567	2.500	-0,067	-2,7

Die Simulationsergebnisse zeigen eine gute Übereinstimmung mit den Ergebnissen, die die Wartenschlangentheorie liefert. Aus den geringen Abweichungen zwischen den entsprechenden Werten kann man entnehmen, daß sich die verwendeten Methoden zur Minderung des Einflusses der Autokorrelation (siehe 5.2), insbesondere die Einteilung in Intervalle, bewähren, da die tatsächlich auftretenden Unterschiede innerhalb des Konfidenzbereichs liegen und dieser damit bestätigt wird.

Beispiel 2:

Am Modell zu Beispiel 1 wird folgende Änderung vorgenommen: Alle Bedienungszeiten sind Erlang-5- verteilt. Die sonstigen Verteilungsparameter aus Beispiel 1 werden beibehalten.
Für das mathematische Modell wird angenommen, daß die Kanäle Bedienungssysteme vom Typ $M/E_5/1$ sind.

a) Wartezeiten an den Kanälen: $EW(WZ)=0{,}2500$ (Math. Modell)

Kanal	$\bar{x}_{WZ}$	Varianz	$EW(WZ)-\bar{x}_{WZ}$	$\frac{EW(WZ)-\bar{x}_{WZ}}{\bar{x}_{WZ}}$ (%)
K_{12}	0,2614	0,0763	-0,0114	- 4,3
K_{23}	0,1720	0,0568	0,0780	44,9
K_{34}	0,1298	0,0352	0,1202	93,2
K_{45}	0,1222	0,0267	0,1278	104,6
K_{56}	0,0819	0,0111	0,1681	205,2

b) Verzögerungen an den Kanälen: $EW(VZ)$

Kanal	$\bar{x}_{VZ}$	Varianz	$EW(VZ)-\bar{x}_{VZ}$	$\frac{EW(WZ)-\bar{x}_{WZ}}{\bar{x}_{WZ}}$ (%)
K_{12}	0,3434	0,0773	-0,o101	- 2,9
K_{23}	0,2564	0,0587	0,0769	29,9
K_{34}	0,2137	0,0367	0,1196	56,0
K_{45}	0,2057	0,0282	0,1276	61,5
K_{56}	0,1665	0,0122	0,1668	100,0

c) Gesamtwartezeiten und Gesamtverzögerungen auf der Teststrecke $R_1 \rightarrow R_5$

Wartezeiten:

$\bar{x}_w$	EW(WZ)	$EW(WZ)-\bar{x}_{WZ}$	$\frac{EW(WZ)-\bar{x}_{WZ}}{\bar{x}_{WZ}}$ (%)
0,729	1,250	0,520	71,4

Verzögerungen:

$\bar{x}_v$	EW(VZ)	$EW(VZ)-\bar{x}_{VZ}$	$\frac{EW(VZ)-\bar{x}_{VZ}}{\bar{x}_{VZ}}$ (%)
1,147	1,667	0,519	45,2

Bei der Berechnung der Werte mit Hilfe des mathematischen Modells wurden für die Kanäle $M/E_5/1$-Wartesysteme zugrundegelegt. Da der aus dem ersten Kanal K_{12} austretende Strom nicht mehr rekurrent ist, sind die beträchtlichen Abweichungen von den simulierten Werten zu erwarten (vgl. 3.1). Die im Auftragsstrom auftretenden Abhängigkeiten machen sich deutlich bemerkbar und die in Kap. 4 aufgrund theoretischer Betrachtungen getroffene Feststellung, daß die Wartezeiten bzw. Verzögerungen von Stufe zu Stufe kürzer werden müßten, wird ebenfalls bestätigt.

Beispiel 3

Die Struktur des Rechnernetzes ist in Bild 9 dargestellt. Es besteht aus 8 Rechensystemen, die miteinander Nachrichten austauschen. Der Weg einer Nachricht durch das Kommunikationssystem wird durch die Nachfolgematrix N (Routing-Table) bestimmt mit

$$n_{ij} = \{\text{Nummer der nächsten Vermittlungseinrichtung auf derm Weg von } V_i \text{ nach } V_j\}$$

$$N = \begin{pmatrix} - & 2 & 3 & 2 & 3 & 3 & 3 & 3 \\ 1 & - & 1 & 4 & 1 & 4 & 1 & 4 \\ 1 & 1 & - & 5 & 5 & 5 & 5 & 5 \\ 2 & 2 & 2 & - & 6 & 6 & 6 & 6 \\ 3 & 3 & 3 & 6 & - & 6 & 7 & 6 \\ 4 & 4 & 5 & 4 & 5 & - & 5 & 8 \\ 5 & 5 & 5 & 5 & 5 & 5 & - & 5 \\ 6 & 6 & 6 & 6 & 6 & 6 & 6 & - \end{pmatrix}$$

Für die Stärke der Auftragsströme zwischen den Knoten des Netzes soll angenommen werden, daß jeder Knoten an jeden anderen Knoten Nachrichten mit einer Stärke umgekehrt proportional zum Quadrat der Entfernung abgibt (Entfernung = Zahl der Kanäle auf dem Weg). Eine Ausnahme bildet der Nachrichtenstrom von V_2 nach V_7, da der Weg von V_2 nach V_7 als Teststrecke dienen soll, und eine geringe Intensität auf dieser Strecke den Aufwand an Simulationszeit zum Erreichen einer bestimmten Nachrichtenzahl vergrößert. Nimmt man die Ankunftsrate der an jedem Knoten von seinen unmittelbaren Nachbarknoten eintreffenden Nachrichten zu $\lambda = 8$ an, so ergibt sich unter den genannten Bedingungen die Auftragsstrom-Matrix Λ mit

$$\lambda_{ij} = \{\text{Auftragsrate von } V_i \text{ nach } V_j\}$$

$$\Lambda = \begin{pmatrix} 0 & 8 & 8 & 4 & 4 & 2 & 2 & 1 \\ 8 & 0 & 3 & 8 & 2 & 4 & 8 & 2 \\ 8 & 4 & 0 & 2 & 8 & 4 & 4 & 2 \\ 4 & 8 & 2 & 0 & 4 & 8 & 2 & 4 \\ 4 & 2 & 8 & 4 & 0 & 8 & 8 & 4 \\ 2 & 4 & 4 & 8 & 8 & 0 & 4 & 8 \\ 2 & 1 & 4 & 2 & 8 & 4 & 0 & 2 \\ 2 & 2 & 2 & 4 & 4 & 8 & 2 & 0 \end{pmatrix}$$

Die Bedienungszeiten in den Vermittlungseinrichtungen sind Erlang-2-verteilt. Die Bedienungszeit s_K in den Kanälen wird bestimmt von der Kanalkapazität k und den Längen l_i der Nachrichten:

$$s_K = \frac{l_i}{k} \qquad \text{(vgl. 2.1)}$$

Für alle Kanäle gilt: k=50.
Die Verteilungsparameter werden so gewählt, daß für alle Teilsysteme des Netzes die Stationaritätsbedingung $\rho<1$ erfüllt ist: Die Länge der Nachrichten wird bei ihrer Erzeugung aus einer (0,3;1,4) Gleichverteilung bestimmt.
(Sie bleibt während des Transports im Gegensatz zu den Nachrichtenlängen im mathematischen Modell konstant.)
Die Bedienungsraten μ_i in den Vermittlungseinrichtungen werden ebenfalls mit Rücksicht auf die Stationaritätsbedingung wie folgt gewählt:

i	1	2	3	4	5	6	7	8
λ_i	81	74	85	79	118	118	53	44
μ_i	95	95	95	95	140	140	62	62
ρ_i	0,85	0,78	0,89	0,83	0,84	0,84	0,85	0,85

Ergebnisse

EW_1 : Erwartungswert, math. Modell $M/E_2/1$

EW_2 : Erwartungswert, math. Modell $M/M/1$

a) Wartezeiten in den Vermittlungseinrichtungen V_i

i	$\bar{x} \cdot 10^{-1}$	$EW_1 \cdot 10^{-1}$	$\frac{EW_1 - \bar{x}}{\bar{x}}$ (%)	$EW_2 \cdot 10^{-1}$	$\frac{EW_2 - \bar{x}}{\bar{x}}$ (%)
1	0,4551	0,4568	0,4	0,6090	33,8
2	0,3426	0,2782	-14,3	0,3709	14,3
3	1,0410	0,5711	-35,5	0,8947	-14,1
4	0,4324	0,3898	- 9,9	0,5198	20,2
5	0,4711	0,2873	-39,0	0,3831	-18,7
6	0,2599	0,2873	10,6	0,3831	47,4
7	0,7079	0,7124	0,6	0,9498	34,2
8	0,3232	0,2957	-8,3	0,3942	22,3

b) Gesamtwartezeiten auf der Teststrecke $V_2 \rightarrow V_1 \rightarrow V_3 \rightarrow V_5 \rightarrow V_7$

$\bar{x}$	EW_1	$\frac{EW_1 - \bar{x}}{\bar{x}}$ (%)	EW_2	$\frac{EW_2 - \bar{x}}{\bar{x}}$ (%)
2,9997	2,4058	-19,8	3,2075	6,9

In diesem Kommunikationsnetz-Modell tritt zusätzlich zu den durch die Erlang-verteilten Bedienungszeiten bedingten Abhängigkeiten in den Nachrichtenströmen ein weiterer Effekt auf: Die Nachrichten behalten ihre Länge während der gesamten Übertragung bei und damit ist auch die Bedienungszeit für eine bestimmte Nachricht an allen Kanälen konstant. Die Nachrichtenströme innerhalb des Netzes sind dadurch so sehr "gestört", daß das mathematische Modell nur noch grobe Näherungswerte für die Wartezeiten liefert.

6. Schlußbemerkung

Für einzelne (isolierte) Wartesysteme mit Poisson-Ankunftsstrom und rekurrentem Bedienungsprozeß liefert die Warteschlangentheorie brauchbare Ergebnisse. Betrachtet man hingegen Serienschaltungen oder Netwerke von Warteschlangen, so zeigt sich, daß eine math. Modellbildung bzw. Analyse solcher Systeme nur unter ganz bestimmten Voraussetzungen, wie z.B. exponentiell verteilte Bedienungszeiten, möglich ist. Da Serienschaltungen und Netze von Warteschlangen in Rechnernetz-Kommunikationssystemen naturgemäß eine wichtige Rolle spielen, wurde auf den warteschlangentheoretischen Hintergrund in dieser Arbeit besonders ausführlich eingegangen. Die Voraussetzungen für eine mathematische Modellbildung, die andererseits wieder eben die Restriktionen bilden, die die Realitätsnähe eines Modells in Frage stellen können, wurden von diesem Hintergrund deutlich gemacht und diskutiert. Die abschließend vorgestellten Simulationsexperimente bzw. die Vergleiche der Ergebnisse des mathematischen Modells mit denen des Simulationsmodells zeigten, daß für eine verfeinerte, realitätsnähere Modellbildung komplexer Warteschlangensysteme, wie sie Rechnernetz-Kommunikationssysteme i.a. darstellen, der Simulation der Vorzug zu geben ist, da mathematische Modelle nur noch grobe Näherungswerte liefern können. Aus diesem Grund wurde auch den Problemen, die bei der Simulation von Warteschlangen auftreten, breiter Raum gewidmet.

Ich schließe mit meinem besonderen Dank an die Herren Holler, Drobnik und Schumacher, auf deren Arbeiten auf dem Gebiet der Rechnernetze ich hier mehrfach verweisen konnte, und die als stets unermüdliche Diskussionspartner zur Klärung der hier vorgestellten Probleme sowie zu ihrer Formulierung beigetragen haben.

7. Anhang: Einige Begriffe der Warteschlangentheorie [19]

Es lassen sich drei Komponenten eines einfachen Wartesystems unterscheiden (siehe Bild 1):

- Auftragsströme
- Warteschlangen
- Bedienungsschalter

Die Kenngrößen eines solchen Systems sind:

λ Ankunftsrate , $\frac{1}{\lambda}$ mittlere Zwischenankunftszeit

μ Bedienungsrate, $\frac{1}{\mu}$ mittlere Bedienungszeit

Die Zwischenankunftszeit Z und die Bedienungszeit S sind i.a. Zufallsvariable, ebenso die Wartezeit WZ und die Verzögerung VZ (VZ=WZ+S).

Bei der Behandlung von Bedienungsproblemen sieht man die Verteilungsfunktionen von Z und S als gegeben an und versucht, weitere interessierende Größen, wie z.B. WZ und VZ zu berechnen. Bei der Anwendung auf reale Bedienungsvorgänge kann diese Methode, wie jede Modellbildung, nur eine Approximation darstellen, da die angenommenen, idealen Eingangsgrößen sich von den wirklichen Eingangsgrößen unterscheiden.

Zur Klassifikation von Warteschlangenmodellen ist die Schreibweise

A/B/c

üblich.

Dabei bedeuten:

A Verteilung der Zwischenankunftszeit Z
B Verteilung der Bedienungszeit S
c Zahl der (parallelen) Bedienungsschalter

Für Verteilungen sind folgende Bezeichnungen gebräuchlich:

M: Exponentialverteilung
E_k: Erlangverteilung mit Parameter k
GI: die Zwischenankunfts- bzw. Bedienungszeiten sind unabhängig identisch verteilte Zufallsgrößen (rekurrenter Prozeß)
G: allgemeine Verteilung
D: Einpunktverteilung

Weitere Unterscheidungen von Wartesystemen sind z.B. hinsichtlich Warteordnung, Warteraum, Abfertigungsmodus usw. möglich. Solange nichts anderes vereinbart ist, wird stets die natürliche Warteordnung (FCFS), unbeschränkter Warteraum und ein Abfertigungsmodus der Art, daß eine begonnene Bedienung ohne Unterbrechung zu Ende geführt werden muß, vorausgesetzt.

Die Folge $\{t_i\}$ der Zeitpunkt des Eintreffens von Aufträgen am Bedienungssystem bildet den Ankunftsstrom ($i = 1, 2, \ldots$). Man kann ihn charakterisieren durch $\tau_i = t_i - t_{i-1}$, also die Zeitspanne zwischen den Ankünften zweier unmittelbar aufeinanderfolgender Aufträge, oder durch die Zahl der Aufträge, die pro Zeiteinheit eintreffen. Werden die τ_i durch Zufallszahlen realisiert, so spricht man von einem zufälligen Strom. Sind sie zudem <u>unabhängig</u> identisch verteilt, gilt also für alle Zwischenankunftszeiten $P(Z \leq z) = F(z)$ ($F(z)$: Verteilungsfunktion), so heißt der Strom <u>rekurrent</u>. In der Literatur findet man dafür auch die Bezeichnung "Strom mit begrenzter Nachwirkung".
Bei der Behandlung von Wartezeitproblemen wird als Eingangsstrom sehr oft ein Spezialfall des rekurrenten Stroms, nämlich der sogenannte <u>Poissonstrom</u>, angenommen.

Dieser wird beschrieben durch:

$$P\{(\text{Zahl der Ankünfte pro Zeiteinheit } t)=k\} =$$

$$=\frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad \lambda > 0$$

Bei der Betrachtung von Warteschlangenmodellen unterscheidet man das transiente (zeitabhängige) und das stationäre (zeitunabhängige) Verhalten des Systems [15,19]. Die Beschränkung auf den stationären Fall, von dem in diesem Bericht ausschließlich die Rede ist, erleichert das Lösen der Modellgleichungen beträchtlich. Stationarität im strengen Sinn bedeutet, daß sämtliche statistischen Eigenschaften des stochastischen Prozesses, den das Modell beschreibt, zeitunabhängig sind. Das Verhalten von Simulationsmodellen kann durch geeignete Tests (z.B. Trend-Test) auf Stationarität geprüft werden, wobei man sich meistens auf den Nachweis der sogenannten schwachen Stationarität, d.h. die Stationarität der Mittelwerte und Autokovarianzen beschränkt, da diese i.a. die Stationarität im strengen Sinn impliziert. Die Stationaritätsbedingung für Wartesysteme lautet:

$$\frac{\lambda}{\mu} < 1$$

Für einige Wartesysteme gibt es im stationären Fall einfach gebaute Formeln zur Bestimmung von Erwartungswert und Varianz der Warte- bzw. Verweilzeit [19,21].

Es bedeuten:

λ	Ankunftsrate
μ	Bedienungsrate
σ_Z^2	Varianz der Zwischenankunftszeit
σ_S^2	Varianz der Bedienungszeit

M/M/1

$$E(WZ) = \frac{\lambda}{\mu(\mu-\lambda)}$$

$$Var(WZ) = \frac{1}{(\mu-\lambda)^2} - \frac{1}{\mu^2}$$

$$E(VZ) = \frac{1}{\mu-\lambda}$$

$$Var(WZ) = \frac{1}{(\mu-\lambda)^2}$$

$M/E_k/1$

$$EW(WZ) = \frac{(k+1)\lambda}{2k\mu(\mu-\lambda)}$$

$$EW(VZ) = \frac{1}{\mu-\lambda}\left(1-\frac{\lambda}{2\mu}\left(1-\frac{1}{k}\right)\right)$$

$$Var(VZ) = \frac{1}{(\mu-\lambda)^2}\left(1-\frac{\lambda}{6\mu}\left(4-\frac{\lambda}{\mu}\right)\left(1-\frac{1}{k}\right)\left(1+\frac{1}{k}\right) - \left(1-\frac{\lambda}{2\mu}\left(1-\frac{1}{k}\right)\right)^2\right)$$

M/G/1

$$EW(WZ) = \frac{\sigma_s^2 + \frac{1}{\mu^2}}{2\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)}$$ (Formel von Pollaczek)

GI/G/1

$$E(WZ) \leq \frac{\sigma_s^2 + \sigma_z^2}{2\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)}$$ (Formel von Kingman)

Literaturverzeichnis

[1] Burke, P.J.
"The Output of a Queuing System", Operations Research 4, 1959, S. 699-704

[2] Burke, P.J.
"The Dependence of Delays in Tandem Queues", Ann. Math. Statist. 35, 1964, S. 874-875

[3] Burke, P.J.
"The Output Process of a Stationary M/M/s Queuing System", Ann. Math. Statist. 39, 1968, S. 1144-1152

[4] Burke, P.J.
"The Dependence of Sojourn Times in Tandem M/M/s Queues", Operations Research 17, 1969, S. 754-755

[5] Conway, R.W. u.a.
"Theory of Scheduling", Addison-Wesley, Reading, Mass., 1967

[6] Conway, R.W.
"Some Tactical Problems in Digital Simulation", Management Science 10, 1, Oktober 1963
S. 47-61

[7] Drobnik, O., Holler, E., Schumacher, F.
"Statusbeschreibung und Statusüberwachung in einem Rechnernetz"
KFK-Ext. 13/72-5, November 1972

[8] Drobnik, O., Holler, E., Schumacher, F.
"Auftragsvergabe in einem Rechnerverbund"
KFK-Ext. 13/72-6, Februar 1973

[9] Finch, P.D.
"The Output of the Queueing System M/G/1",
J.R. Statist. Soc. B 21, 2, 1959, S. 375-380

[10] Friedmann, H.D.
"Reduction Methods for Tandem Queueing Systems",
Operations Research 13, 1, Januar 1965, S. 121-131

[11] Gnedenko, B.W., Kowalenko, I.N.
"Einführung in die Bedienungstheorie", Oldenbourg, München 1971

[12] Jackson, J.R.
"Networks of Waiting Lines", Operations Research 5,
1957, S. 518-521

[13] Kingman, J.F.C.
"Inequalities in the Theory of Queues", J. Roy.
Statist. Soc. B, 32, 1, 1970, S. 102-110

[14] Kleinrock, L.
"Communication Nets, Stochastic Message Flow and
Delay",Mac Graw-Hill, New York, 1964

[15] Mihram, G.A.
"Simulation, Statistical Foundations and Methodology",
Academic Press, New York, 1972

[16] Muntz, R.R.
"Poisson Departure Processes and Queueing Networks"
IBM Research Report RG 4145, Dezember 1972

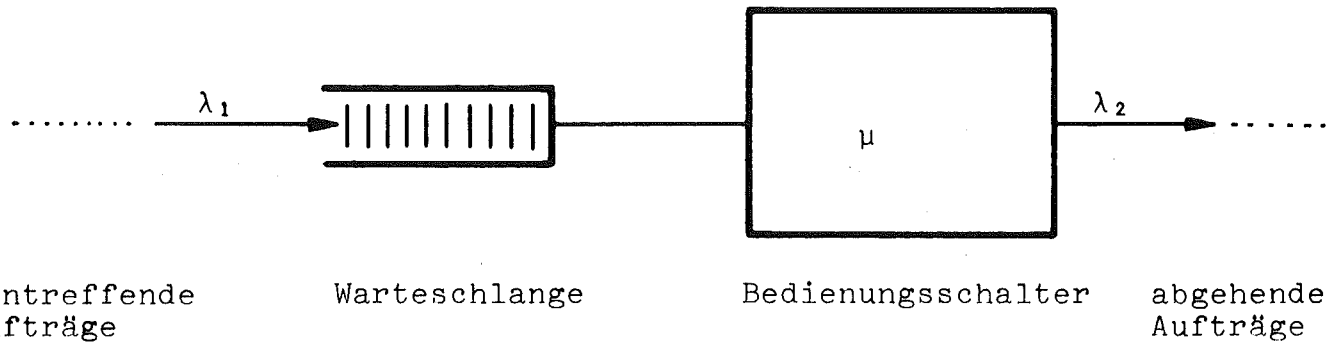
[17] Reich, E.
"Waiting Times when Queues are in Tandem",
Operations Research 5, 1957, S. 768-773

[18] Reich, E.
"Note on Queues in Tandem", Ann. Math. Statist. 34,
1963, S. 338-341

[19] Saaty, T.L.
"Elements of Queueing Theory", Mc Graw-Hill,
New York, 1961

[20] Senger, R.
"Auftrags-Wartezeiten als Maß für die Bedienungsqualität
eines Rechnerverbundnetzes",
Diplomarbeit, Universität Karlsruhe, 1972

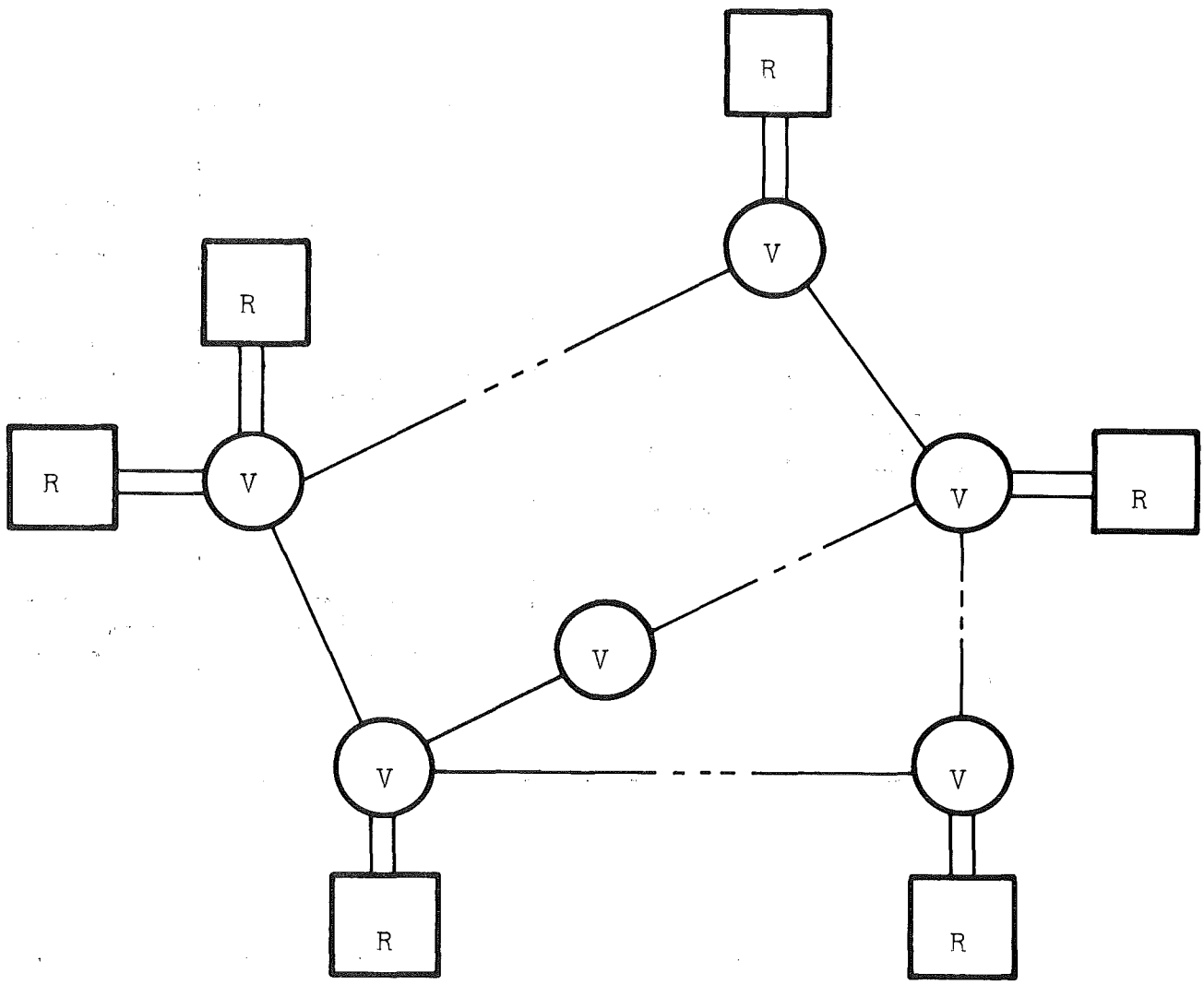
[21] -.-
"Analysis of Some Queueing Models in Real-Time Systems"
IBM GF 20-0007-1



λ_1, λ_2 : Ankunfts- bzw. Abgangsrate

μ : Bedienungsrate

Bild 1: Einfaches Bedienungssystem



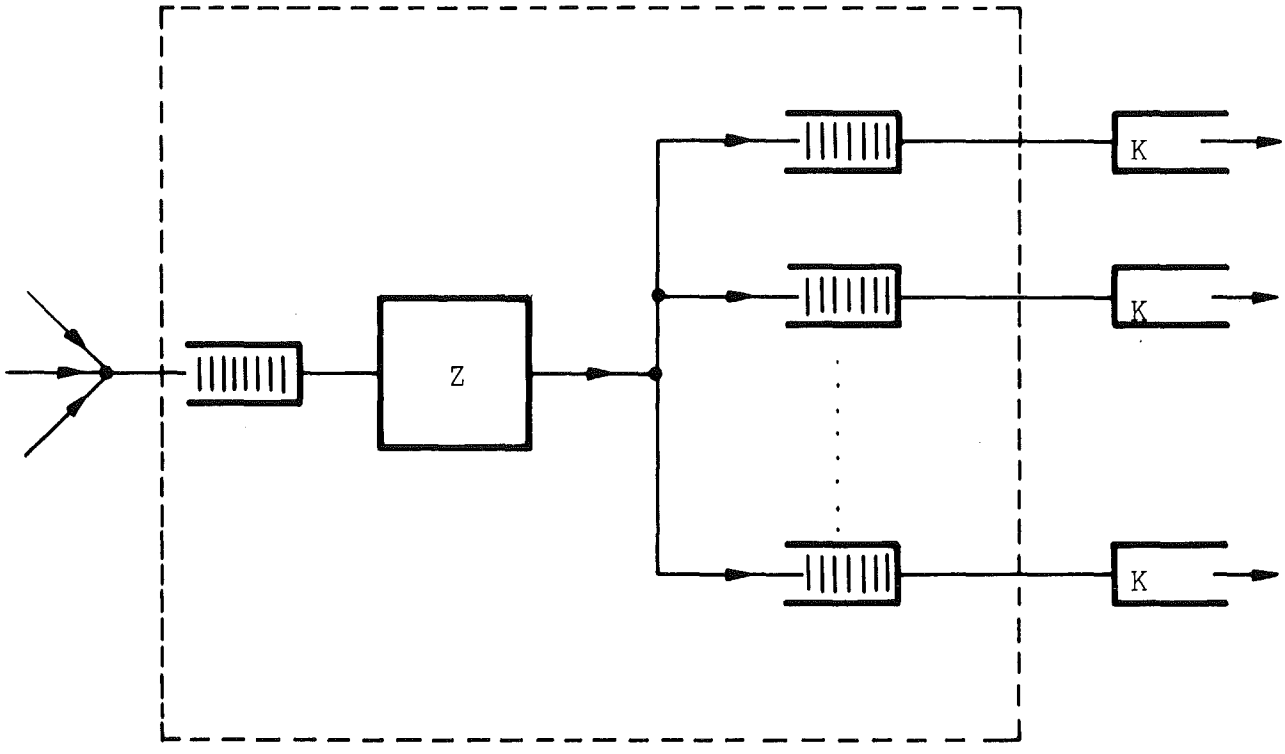
R : Rechner

V : Vermittlungseinrichtung

═══ Kopplungskanal zwischen Rechner und Vermittlungseinrichtung

——— Übertragungskanal zwischen Vermittlungsknoten

Bild 2: Rechner und Vermittlungseinrichtungen im Rechnernetz



K : Kanal

Z : Zuweisung einer eingehenden Nachricht an eine Kanalwarteschlange

Bild 3: Warteschlangenmodell eines Vermittlungsknotens bei Message-Switching

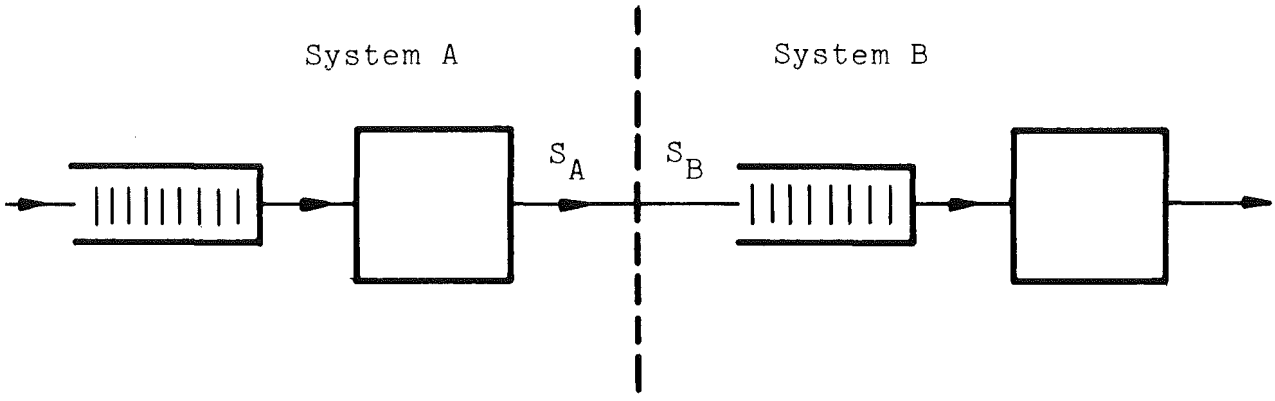
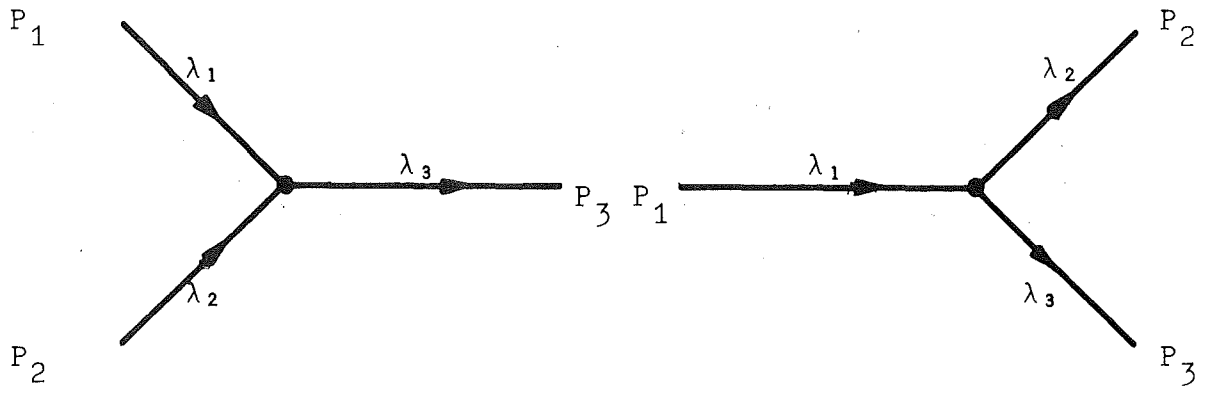


Bild 4: Serienschaltung von zwei Bedienungssystemen



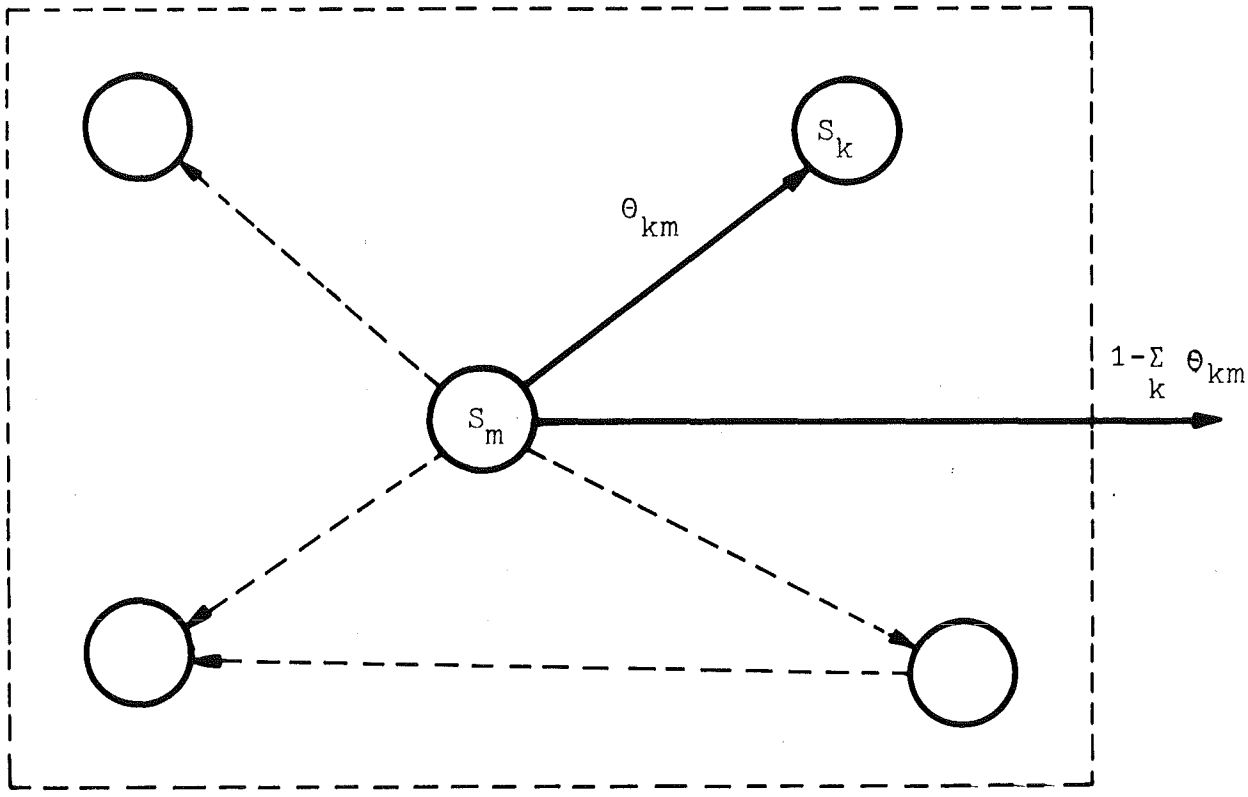
P_1, P_2, P_3 Poissonströme
λ_1, λ_2, λ_3 Ankunftsraten

a) Überlagerung: $\lambda_3 = \lambda_1 + \lambda_2$

b) Aufspaltung: $\lambda_2 = r\lambda_1$
$\lambda_3 = (1-r)\lambda_1$

r=P (Auftrag aus P_1 setzt seinen Weg in P_2 fort)

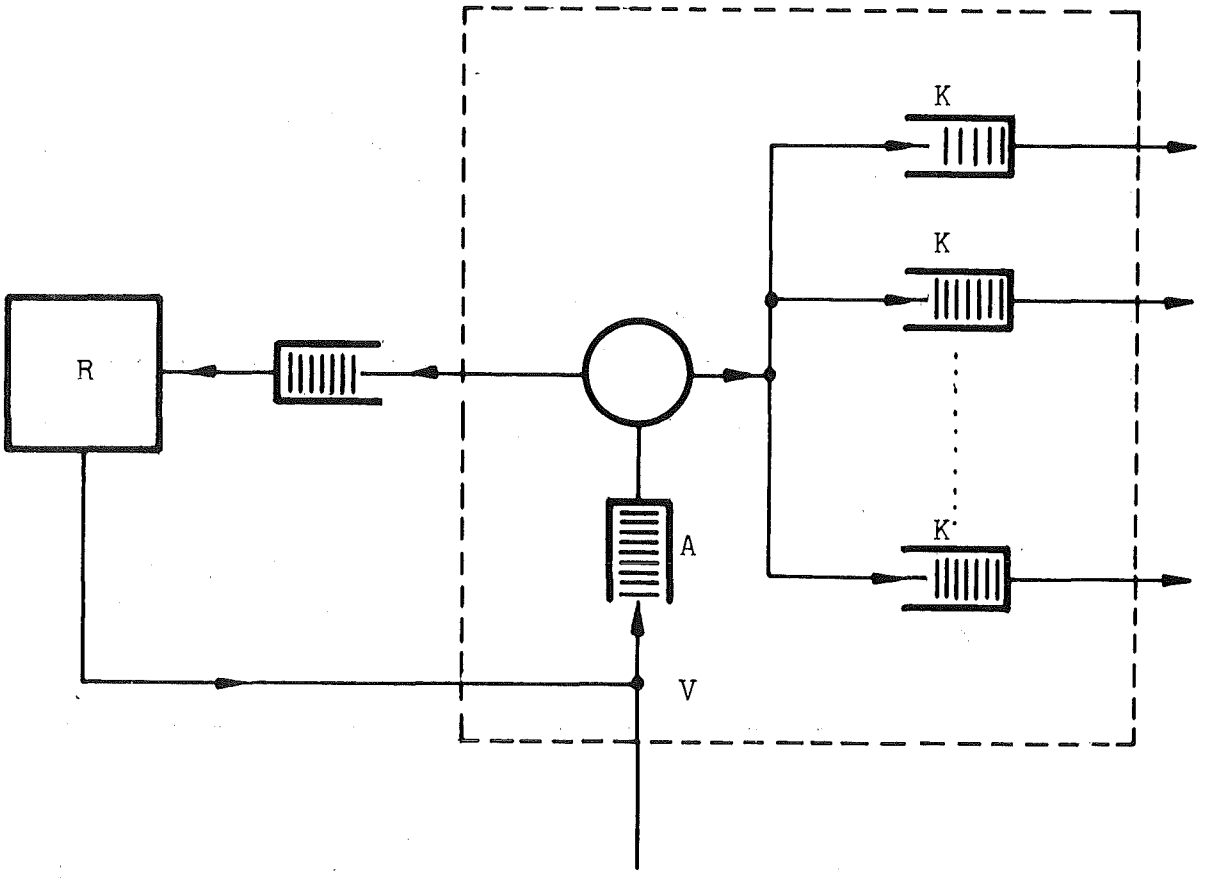
Bild 5: Überlagerung und Aufspaltung von Poissonströmen



S_m, S_k Bedienungssysteme mit negativ expontentiell verteilter Bedienungszeit

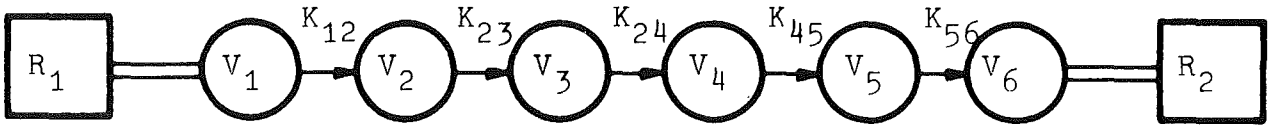
Θ_{km} = P (Auftrag, der in S_m abgefertigt wurde, geht nach S_k)

Bild 6: Netz aus Bedienungsstationen, in dem sich die Teilsysteme wie unabhängige Systeme verhalten



R: Rechner
V: Vermittlungseinrichtung
K: Kanalwarteschlangen für ausgehende Nachrichten
A: Warteschlange für die an der Vermittlungseinrichtung eintreffenden Nachrichten

Bild 7: Umgebung eines Rechners im Simulationsmodell



R_i : Rechner
V_i : Vermittlungseinrichtungen
K_{ij} : Übertragungskanäle zwischen den Vermittlungseinrichtungen

Bild 8: Nachrichtenübermittlung zwischen zwei Rechnern
(Simulation, Beispiel 1 und 2, siehe 5.3)

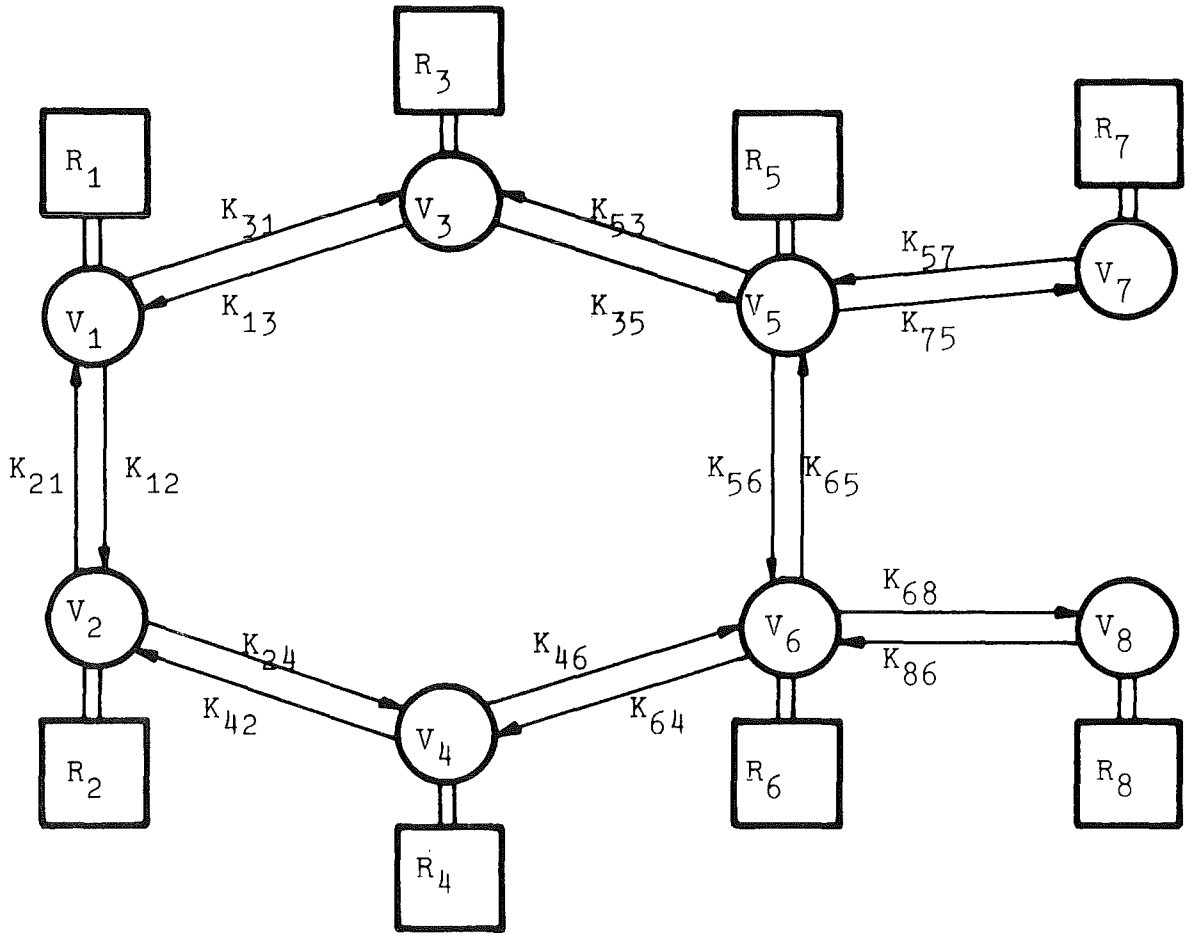


Bild 9: Struktur des in Beispiel 3 (5.3) simulierten Rechnernetzes