GI'79, Informatik Fachberichte 19, Springer Verlag Berlin

ZUR ANWENDUNG OPTIMALER BEDIENSTRATEGIEN IN RECHNERN*

Bernhard Walke

AEG-TELEFUNKEN
Forschungsinstitut Ulm

7900 Ulm/Donau

Zusammenfassung: Es wird die Anwendung eines bekannten Kriteriums [1,2] zur kostenminimalen Bedienung wartender Aufgaben anhand von Beispielen erläutert. Das Kriterium ist vor allem bei stark streuenden Bediendauerverteilungen der Aufgaben nützlich. Für diesen Fall wird gezeigt wie -durch geschickte Modellierung- eine einfache praktikable Bedienstrategie entsteht, deren mittlere Wartezeit analytisch berechenbar ist.

1. EINLEITUNG

Bei der rechnerischen Behandlung von Verkehrsmodellen zur Nachbildung von Betriebsmitteln (z.B. Rechnerkern, Transportkanal, Hauptspeicher, usw.) einer Rechenanlage kann man zwei Aufgabenstellungen unterscheiden:

- die Berechnung charakteristischer Größen, wie Durchsatz, Wartezeitverteilung (oder deren Momente), Warteschlangenlänge, usw. bei gegebener Bedienungsstrategie; diese Aufgabe wird Analyse genannt.
- die Bestimmung der Bedienungsstrategie so, daß eine betrachtete charakteristische Größe optimiert (z.B. die mittlere Wartezeit minimiert) wird; diese Aufgabe wird Synthese genannt.

Während die erste Aufgabe mit Methoden der Warteschlangen-Theorie (queueing theory) gelöst wird, verwendet man zur Lösung der zweiten Aufgabe die Hilfsmittel der Bedienungs-Theorie (scheduling theory).

Im Themenkreis der Bedienungstheorie kann man wieder zwei Schwerpunkte bilden. Sie behandeln

- die optimale Ablauforganisation mit vorausbekannter Last, z.B. die Bearbeitung von n vorliegenden Aufgaben auf m Verarbeitungseinheiten in möglichst kurzer Zeit, wobei die Bediendauern der Aufgaben bekannt sind (job shop problem),
- den stochastischen Betriebsfall mit zufällig ankommenden Aufgaben, deren Bediendauern bekannten Wahrscheinlichkeitsverteilungen gehorchen.

---

* Die Arbeiten wurden zu 50 % durch das 3. DV-Programm der Deutschen Bundesregierung FKZ 081 2248 gefördert.

Im folgenden beschäftigen wir uns ausschließlich mit dem letztgenannten Fall. Es soll die Anwendung eines fundamentalen Ergebnisses von Olivier [1,2] anhand von Beispielen gezeigt werden, wie sie in der praktischen Anwendung bei Rechenanlagen im Echtzeitbetrieb auftreten. Eingeschlossen sind dabei sowohl Teilnehmer- wie Prozeßrechensysteme.

2. DAS MODELL

Wir betrachten zunächst ein Modell eines Rechners mit einem Poisson-Ankunftsprozeß mit der Ankunftsrate λ [Aufgaben pro Zeiteinheit] entsprechend Bild 1. Alle ankommenden Aufgaben

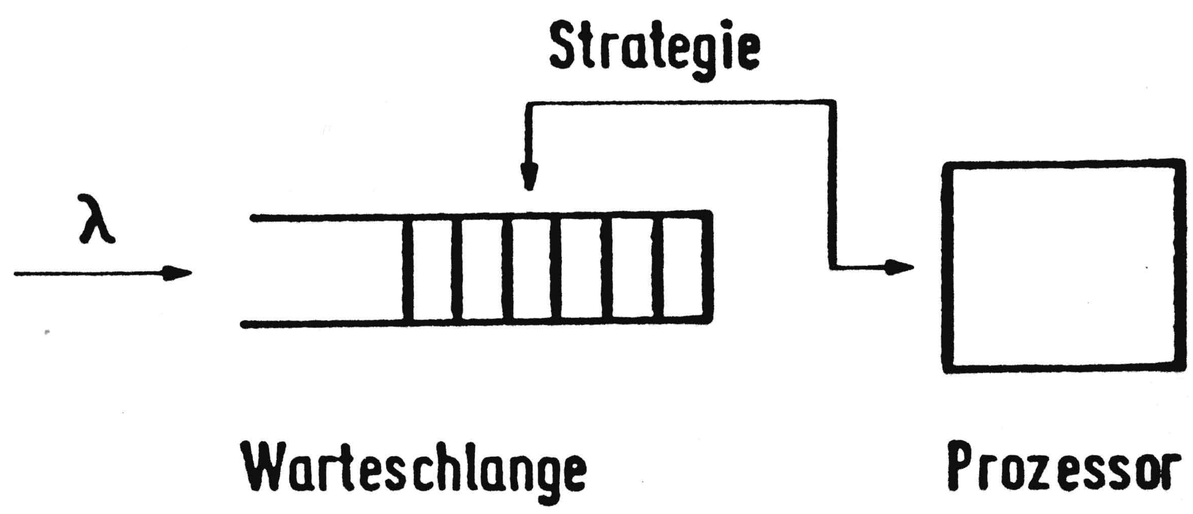


Bild 1: Modell vom Typ M/G/1 zur Nachbildung der Bedienung von Aufgaben eines Rechners. λ : Ankunftsrate pro Zeiteinheit (= Durchsatz).

warten in dem als Warteschlange dargestellten Warteraum, bis sie fertig bearbeitet sind. Die Bediendauern b aller Anforderungen gehorchen einer Verteilungsfunktion $B(t) = P(b \leq t)$. Gesteuert durch eine Bedienungsstrategie wählt der Prozessor eine der wartenden Aufgaben aus dem Warteraum zur Bearbeitung aus und bedient sie solange bis

- sie fertiggestellt ist oder
- die Bedienungsstrategie diese Bearbeitung abbricht - was zu einer Verdrängung der Aufgabe aus dem Prozessor zurück in den Warteraum führt - um eine andere Aufgabe zu bearbeiten.

Eine in ihrer Bearbeitung unterbrochene Aufgabe wird später fortgesetzt, ohne daß schon geleistete Prozessorarbeit wiederholt wird. In einer für Wartemodelle üblichen Abkürzungsform [5,Kap.1.5] handelt es sich um ein Modell des Typs M/G/1.

Gesucht ist eine Bedienungsstrategie, bei der die erwartete Wartezeit W aller Aufgaben minimal wird. Die Wartezeit w ist definiert

w = Fertigstellungs- minus Ankunftszeitpunkt
minus Bediendauer.

Es ist $W = E\{w\}$ der Erwartungswert der Wartezeit aller Aufgaben.

Das Modell nach Bild 1 läßt sich im Rahmen dieser Fragestellung so verallgemeinern, daß N verschiedene Typen von Aufgaben nebeneinander erlaubt sind, deren gemeinsame erwartete Wartezeit minimiert werden soll, Bild 2.

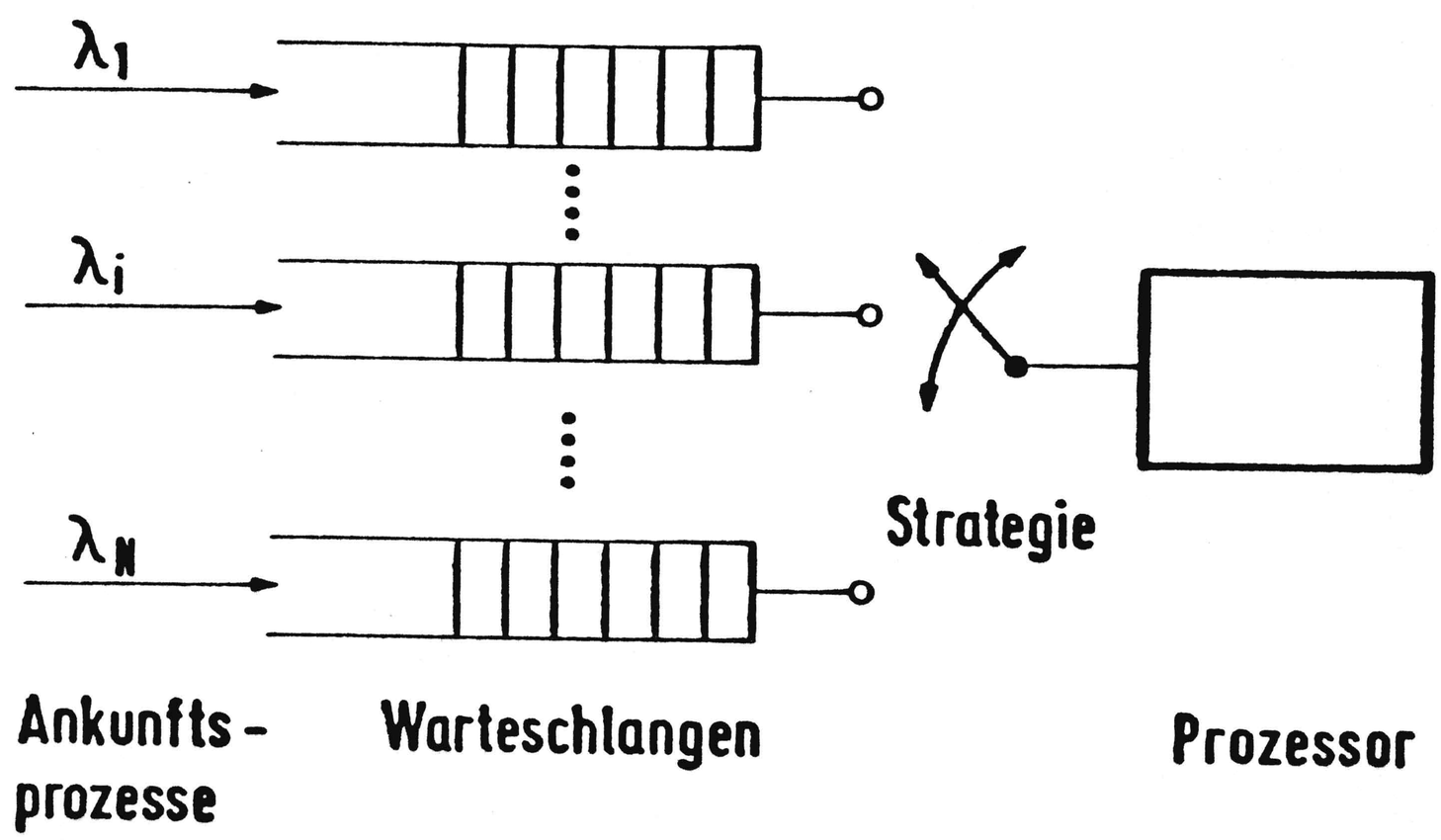


Bild 2: Modell eines Rechners mit -im Unterschied zu Bild 1- N Typen von Aufgaben mit je individueller Bediendauerverteilung.

Die Ankunftsrate von Aufgaben des Typs ist jetzt λ_i und die Bediendauer-Verteilungsfunktion $B_i(t)$. Es werden unabhängige Poisson-Ankunftsprozesse und unabhängige Bediendauerverteilungen angenommen.

Es ist üblich, bei der Optimierung der Bedienung von Aufgaben auch vom Typ abhängige Kostenfunktionen $g_i(t)$ zuzulassen. Eine Aufgabe, die insgesamt die Dauer t durch den Prozessor bedient wurde und bisher die Dauer w_i gewartet hat, hat die Wartekosten $g_i(t)w_i$. Die Wartezeit geht linear in die Kosten ein, während die Kosten beliebig von der tatsächlichen Bediendauer einer individuellen Aufgabe abhängen dürfen. Optimierungsziel ist nun die Minimierung der erwarteten Wartekosten.

$$K = E\left\{\sum_{i=1}^{N} g_i(t)w_i\right\} .$$

Anstelle der Wartezeit w_i könnte man auch die Verweilzeit $v_i = b_i+w_i$ mit Kosten belegen. Dieser Ansatz führt jedoch auf dasselbe Optimierungsziel.

Wegen $K' = E\left\{\sum_{i=1}^{N} g_i(t)[w_i+b_i]\right\} = K + E\left\{\sum_{i=1}^{N} g_i(t)b_i\right\} = K + K''$

enthält K' offenbar einen strategieunabhängigen additiven Anteil K", so daß weiterhin die erwarteten Wartekosten K allein (über die Wartezeit w_i) strategieabhängig sind.

Bild 3 zeigt Beispiele für Kostenfunktionen: In der praktischen An-

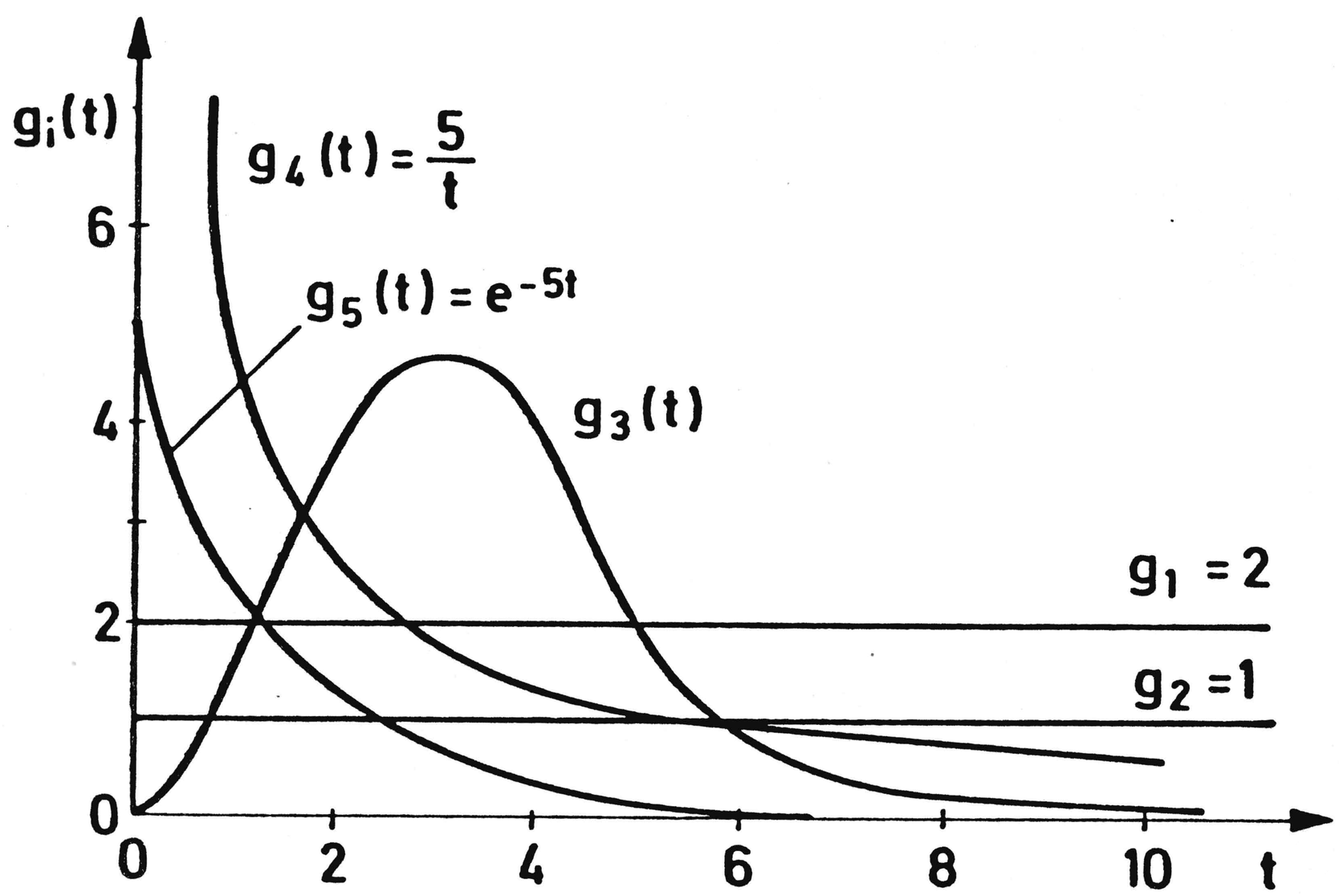


Bild 3: Beispiele für Kostenfunktionen.

wendung haben konstante Kostenfunktionen, z.B. $g_i=a$ (a = natürliche Zahl), die größte Bedeutung. Mit zunehmender Bediendauer t fallende Kostenfunktionen führen zu einer starken Bevorzugung von Aufgaben mit kurzer Bediendauer. Die Kostenfunktion $g_3(t)$ in Bild 3 fördert besonders die Bedienung von Aufgaben, die höchstens 4 Zeiteinheiten Bediendauer haben. Aufgaben mit größeren Bediendauern müssen hier ein Vielfaches an Wartezeit aufweisen, um mit kurzen Aufgaben kon-

kurrieren zu können.

Bild 4 zeigt den typischen Verlauf der Wartezeit-Verteilungsfunktion von Aufgaben für ein Belastungsbeispiel mit zwei Typen i = 1,2 von Aufgaben des Modells nach Bild 2 bei optimaler Bearbeitung. An der Ordinate steht die Wahrscheinlichkeit, daß eine Wartezeit w_i kleiner

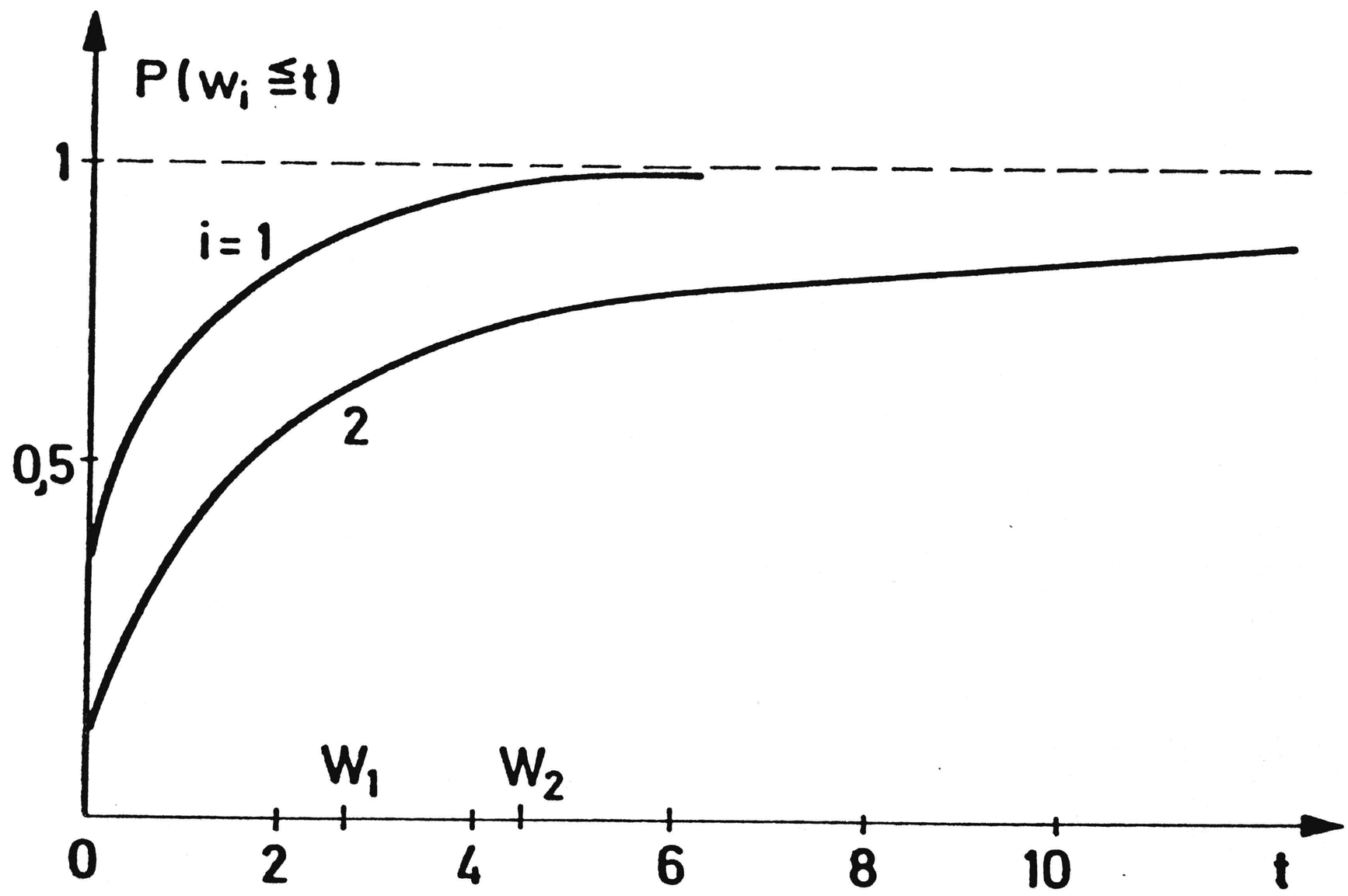


Bild 4: Beispiele für Wartezeit-VFn
W_1, W_2 sind die Erwartungswerte der Wartezeit für Aufgaben des Typs 1 bzw. 2.

oder gleich t ist. Mit hoher Wahrscheinlichkeit sind Wartezeiten unterdurchschnittlich klein, selten überdurchschnittlich groß.

In der praktischen Anwendung ist das Problem zu lösen, bei bekannten Bediendauerverteilungen der Aufgaben Kostenfunktionen so zu wählen, daß ein gewünschter Verlauf der Wartezeitverteilungen resultiert. Alle Kostenfunktionen in Bild 3 haben die Eigenschaft, daß kurze Aufgaben gegenüber längeren bevorzugt werden. Darüber hinaus gehende Forderungen lassen sich nur schwer begründen und dafür passende Kostenfunktionen kaum ermitteln. Deshalb begnügt man sich meist mit konstanten Kostenfunktionen.

Im Modell nach Bild 1 (mit nur einem Typ von Aufgaben) verbleibt dann als Optimierungsziel die minimale erwartete Wartezeit W. Entsprechend ist das Optimierungsziel im Modell nach Bild 2 die Summe

der mit den Kostenfunktionen gewichteten mittleren Wartezeiten

$$K = \sum_{i=1}^{N} g_i W_i \ .$$

3. KRITERIUM FÜR KOSTENMINIMALE BEDIENUNGSSTRATEGIEN

In früheren Arbeiten [1,2] konnte gezeigt werden, daß sich -im Rahmen der betrachteten Modelle- eine kostenminimale Bedienstrategie ergibt, wenn zu jedem Zeitpunkt eine Kenngröße $\varphi^*(i,t)$ für jede wartende Aufgabe des Typs i bestimmt und die Aufgabe mit maximaler Kenngröße bedient wird.

Die Herleitung dieses Verfahrens und der Beweis der Optimalität ist kompliziert. Wir wollen die Anwendung des Ergebnisses besprechen. Dazu wiederholen wir zunächst Ergebnisse aus [2].

Eine Aufgabe mit der Bediendauer-Verteilung $b_i(x) = dB_i(x)/dx$ sei schon die Dauer t bedient worden. In einem zukünftigen Bedieninterval der Länge δ wird diese Aufgabe fertiggestellt mit der Wahrscheinlichkeit

$$\int_t^{t+\delta} \frac{b_i(x)dx}{1-B_i(t)} \ .$$

Unter Verwendung einer allgemeinen Kostenfunktion $g_i(x)$ sind die erwarteten Kosten des Intervalls

$$E\{k_i\} = \int_t^{t+\delta} \frac{g_i(x)b_i(x)dx}{1-B_i(t)} \ . \qquad (3.1)$$

Der Ausdruck $K=g_i(t)b_i(t)/[1-B_i(t)]$ ist die bewertete Erfolgsrate von Aufgaben des Typs i. (3.2)

Der Erwartungswert der effektiven Dauer des Intervalls δ (evtl. wird die Aufgabe ja schon früher fertig) ist

$$E\{\delta\} = [\int_t^{t+\delta}(x-t)b_i(x)dx + \delta\int_{t+\delta}^{\infty} b_i(x)dx]/[1-B_i(t)]. \qquad (3.3)$$

Die Kenngröße $\varphi^*(i,t)$, Leistung genannt, ist

$$\varphi^*(i,t) = \max_{\delta \geq 0} \frac{E\{k_i\}}{E\{\delta\}} . \qquad (3.4)$$

Eine Vereinfachung dieses Ausdruckes ergibt sich, wenn die bewertete Erfolgsrate, Gl.(3.2), monoton mit t steigt. Dann wird

$$\varphi^*(i,t) = \frac{t\int_t^\infty g_i(x)b_i(x)dx}{\int_t^\infty (x-t)b_i(x)dx} . \qquad (3.5)$$

Fällt die bewertete Erfolgsrate monoton mit t, dann wird

$$\varphi^*(i,t) = \frac{g_i(t)b_i(t)}{1-B_i(t)} , \qquad (3.6)$$

also gleich dieser Rate selbst.

Diese Monotonie-Eigenschaften findet man bei den meisten üblichen Bediendauerverteilungen (Bild 5). Z.B. sind für $g_i(t) = g_i$ die Erfolgsraten monoton für negativ exponentielle, hyperexponentielle, entartet exponentielle, n-Punkt, Erlang-k-Verteilungen sowie Sonderfälle anderer Verteilungen (z.B. der stückweise exponentiellen und der allgemeinen Erlangverteilung).

Es könnte nun der Eindruck entstehen, daß die Verwirklichung einer kostenminimalen Bedienstrategie den ständigen Vergleich der Leistungen $\varphi^*(i,t)$ $(i=1,2,\ldots,N)$ aller wartenden Aufgaben und, als Voraussetzung dafür, die ständige Ermittlung von $\varphi^*(i,t)$ erforderlich macht. Die folgenden Beispiele sollen zeigen, daß solche Rechnungen und Vergleiche u.U. nur selten nötig sind. Für pathologische Fälle, bei denen solche Rechnungen ständig erforderlich sind, wird eine Möglichkeit gezeigt, durch eine unbedeutende Änderung der Voraussetzungen die Entscheidungshäufigkeit der Bedienstrategie ebenfalls sehr klein zu machen. Das eben eingeführte Kriterium zur Auswahl einer wartenden Aufgabe zur Bedienung ist unverändert auch anwendbar, wenn anstelle nur eines Prozessors gleichzeitig mehrere parallele Prozessoren vorhanden sind.

4. ANWENDUNGSBEISPIELE MIT KOSTENMINIMALER BEDIENSTRATEGIE

In den folgenden Beispielen nehmen wir konstante Kostenfunktionen $g_i(t) = g_i$ an.

4.1 Erlang-K, negativ exponentielle und konstante Verteilungen

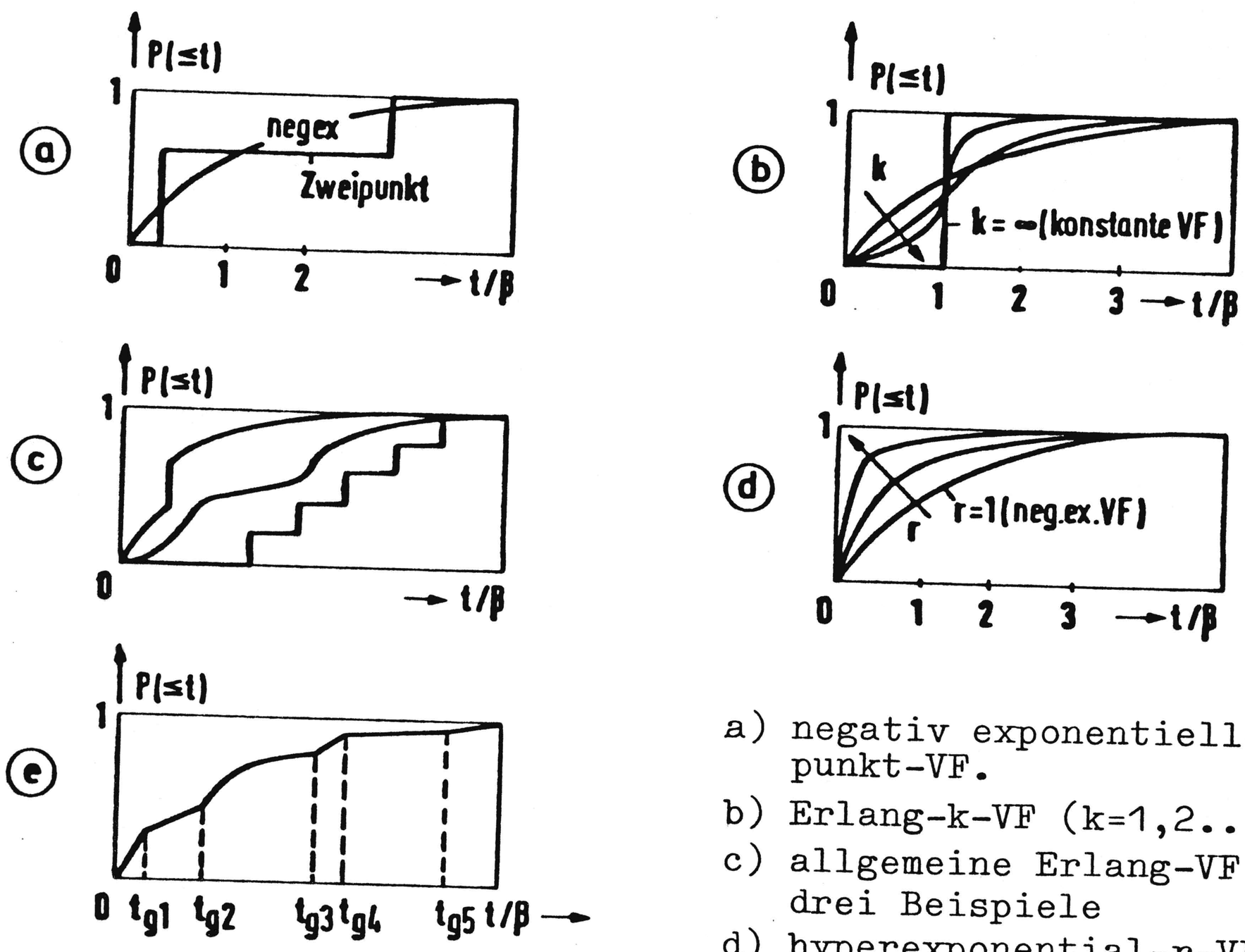
Die Erlang-k-Verteilung (E_k) enthält als Grenzfälle für k=1 die negativ exponentielle (M) und für $k \to \infty$ die konstante (D) Verteilung, vgl. Bild 5a, b. Verteilungsfunktion B(t) und Verteilung b(t) lauten

$$B(t) = 1-\exp(-k\mu t) \sum_{j=o}^{k-1} \frac{(k\mu t)^j}{j!}; \quad b(t) = \frac{(k\mu)^k}{(k-1)!} t^{k-1} \exp(-k\mu t).$$

Für diese Klasse von Verteilungen steigt die bewertete Erfolgsrate nach Gl.(3.2) monoton bzw. ist konstant (für k=1); die Leistung kann nach Gl.(3.5) bestimmt werden zu

$$\varphi^*(i,t) = g_i \frac{k\mu \sum_{j=o}^{k-1} \frac{(k\mu_i)^j}{j!} t^j}{\sum_{j=o}^{k-1} \frac{(k-j)(k\mu_i)^j}{j!} t^j}, \tag{4.1}$$

wobei i der Typ der Aufgabe und t ihre schon verbrauchte Bedienzeit ist.



a) negativ exponentielle, Zweipunkt-VF.
b) Erlang-k-VF (k=1,2...∞).
c) allgemeine Erlang-VF, drei Beispiele
d) hyperexponential-r-VF (r=1,2,... ∞)
e) stückweise exponentielle VF 5-ter Ordnung.

Bild 5: Häufig verwendete Verteilungsfunktionen für Prozesse verkehrstheoretischer Modelle.

Warten nur Aufgaben mit derselben E_k, M oder D Bediendauerverteilung, dann muß die Bedienstrategie unter allen wartenden Aufgaben, diejenige mit größter Leistung auswählen und ihr eine unbegrenzte Rechenfrist zuteilen. Die Leistung einer in Bedienung befindlichen Aufgabe wächst monoton, kann also von keiner der anderen wartenden Aufgaben übertroffen werden.

Unbearbeitete Aufgaben haben gleiche Leistung $\varphi^*(i,0)=\mu_i$ und sind deshalb gleichberechtigt. Beispielsweise ist die Bediendauer in der Reihenfolge der Ankunft (first come first serve, FCFS) optimal. Entscheidungszeitpunkte der Bedienstrategie sind allein die Fertigstellungszeitpunkte von Aufgaben.

Die Leistung von Aufgaben mit sehr großer Bediendauer strebt asymptotisch gegen den Wert $\varphi^*(i,t)=k\mu_i$.
$t \to \infty$

Für interessante Fälle der E_k-Verteilung ergibt sich

E_2-Verteilung: $$\varphi^*(i,t) = \frac{g_i\mu_i(1+2\,\mu_i\,t)}{1+\mu_i t}$$

M-Verteilung: (d.h. k=1) $$\varphi^*(i,t) = \mu_i .$$

D-Verteilung: (d.h. $k \longrightarrow \infty$) $$\varphi^*(i,t) = \frac{g_i\mu_i}{1-\mu_i t} .$$

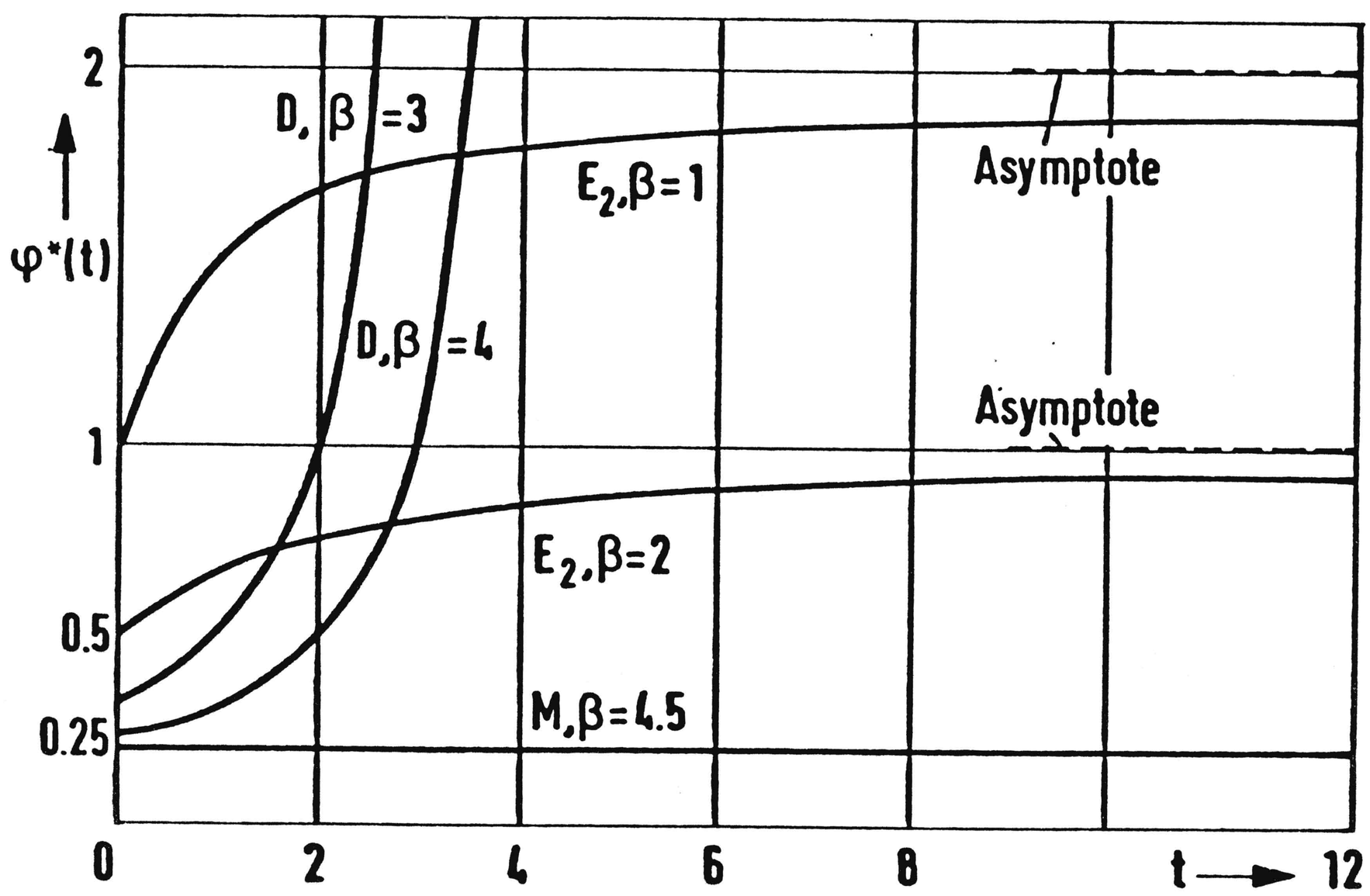


Bild 6: Maximale Leistung für Gewichtsfaktoren $g_i=1$ bei drei Bediendauer-VFn D, M und E_2 in Abhängigkeit von der schon verbrauchten Bearbeitungszeit t. Parameter ist $\beta = 1/\mu$.

Bild 6 zeigt diese Leistungen in Abhängigkeit der verbrauchten Bediendauer t für beispielhaft gewählte Verteilungsparameter.

Für Aufgaben mit E_2-Verteilung überschreitet die Leistung niemals den durch die Asymptote gegebenen Wert. Aufgaben mit M-Verteilung haben eine konstante Leistung (unbearbeitete und bearbeitete Aufgaben sind gleichberechtigt). Die Leistung von Aufgaben mit konstanter Bediendauer steigt umso steiler, je näher man der Fertigstellung kommt und ist unmittelbar vor Fertigstellung unbegrenzt groß.

Konkurrieren Aufgaben mit <u>unterschiedlichen</u> Bediendauer-Verteilungen, z.B. wie in Bild 6 angenommen, dann sind neben den Fertigstellungszeitpunkten auch die Ankunftszeitpunkte Entscheidungszeitpunkte der Bedienstrategie. Es ist nämlich dann möglich, daß eine ankommende Aufgabe einem Typ j angehört, der bisher nicht wartete und eine besonders hohe Startleistung $\varphi^*(j,0)$ besitzt:

Eine unbearbeitete Aufgabe i, die zum Zeitpunkt t_o wegen ihrer hohen Leistung $\varphi^*(i,0)$ in Arbeit genommen wurde, hat eine konstante oder anmonoton steigende Leistung. Sie kann jedoch von einer zur Zeit t_o+t kommenden Aufgabe j mit $\varphi^*(j,0) > \varphi^*(i,t)$ unterbrochen werden.

Im Beispiel nach Bild 6 unterbricht eine ankommende Aufgabe vom Typ $(E_2, \beta=1)$ eine andere vom Typ $D, \beta=3)$, solange letztere noch nicht die Dauer t=2.0 Zeiteinheiten bedient wurde. Andernfalls findet keine Unterbrechung statt. Für eine Unterbrechungsentscheidung ist es ohne Belang, wie sich die Leistung der unterbrochenen Aufgabe, verglichen it der der unterbrechenden, bei weiterer Bedienung entwickelt.

4.2 <u>Hyperexponentielle Verteilungen</u>

Die Hyperexponential-r (H_r-) Verteilung enthält als Grenzfall für r=1 die negativ exponentielle (M-) Verteilung, vgl. Bild 5d. Verteilungsfunktion und Dichte lauten

$$B(t) = 1-\sum_{j=1}^{r} p_j \exp(-\mu_j t); \qquad b(t) = \sum_{j=1}^{r} p_j \mu_j \exp(-\mu_j t) \qquad \text{mit}$$

$$\sum_{j=1}^{r} p_j = 1.$$

Für Verteilungen mit r > 1 <u>fällt</u> die bewertete Erfolgsrate, Gl.(3.2), monoton mit der aufgenommenen Bediendauer; die Leistung kann nach Gl.(3.6) bestimmt werden zu

$$\varphi^*(i,t) = g_i \frac{\sum_{j=1}^{r} p_{ij}\,\mu_{ij}\,\exp(-\mu_{ij}t)}{\sum_{j=1}^{r} p_{ij}\,\exp(-\mu_{ij}t)}, \qquad (4.2)$$

wobei i der Typ der Aufgabe und t ihre verbrauchte Bedienzeit ist. Für unbediente Aufgaben ist $\varphi^*(i,t) = \sum_{j=1}^{r} p_{ij}\,\mu_{ij}$, für $t \to \infty$ erhält man $\lim_{t\to\infty} \varphi^*(i,t) = \mu_{ir}$, wobei μ_{ir} der kleinste Wert unter allen μ_{ij} ist.

Bild 7 zeigt ein Beispiel für zwei Typen von Aufgaben i=1,2, die beide H_2-Bediendauerverteilungen (aber verschiedene Parameter μ_{ij}) haben. Konkurrieren mehrere Aufgaben desselben Typs um den Prozessor, dann gibt es Situationen, in denen zwei oder mehr Aufgaben dieselbe Bedienzeit t aufweisen und gleiche Leistung haben.

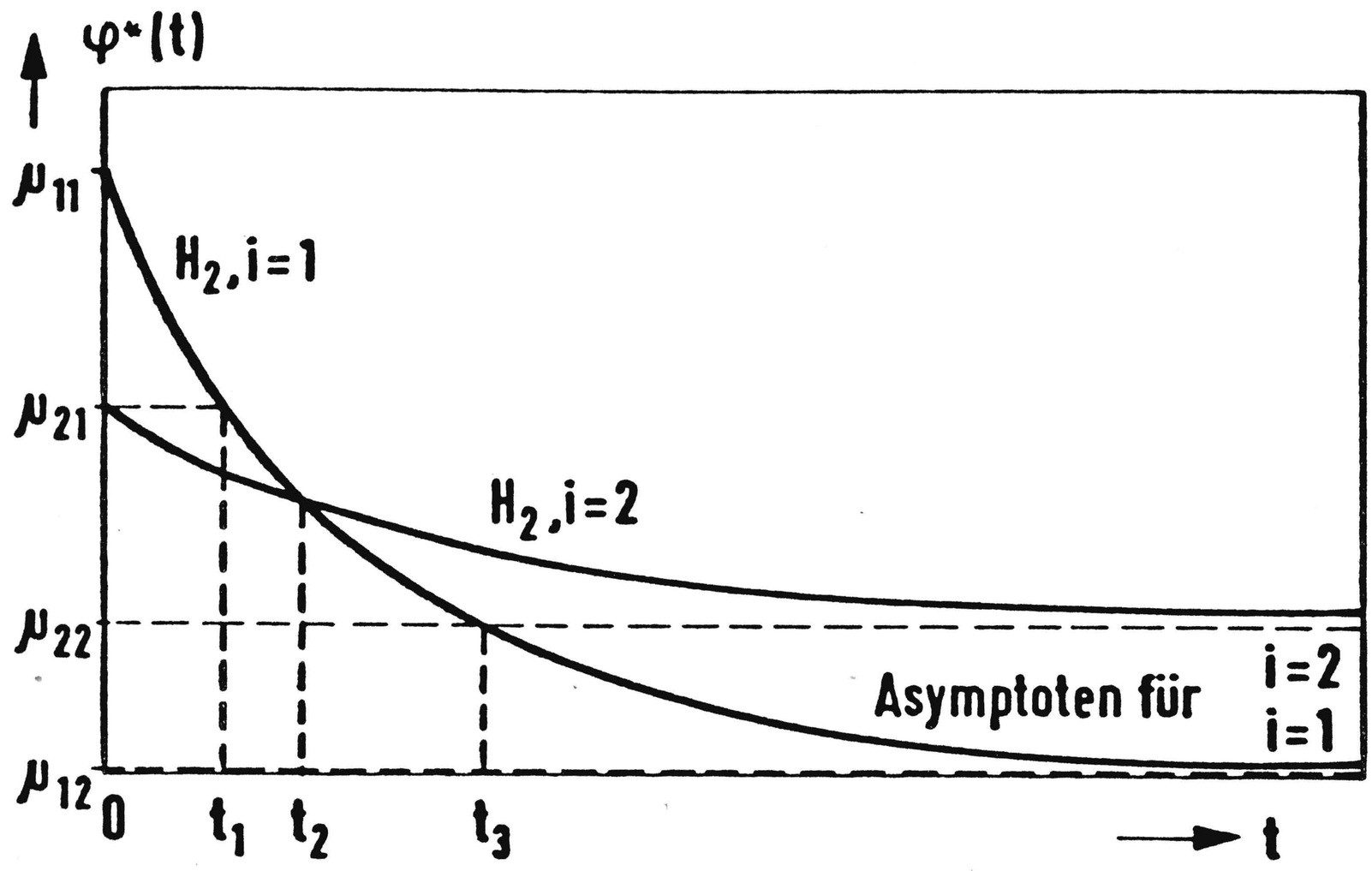


Bild 7: Maximale Leistung für Gewichtsfaktoren $g_i=1$ für zwei Typen i=1,2 von Aufgaben mit unterschiedlichen H_2-Bediendauer-VFn. in Abhängigkeit von der verbrauchten Prozessorzeit t.

Sobald eine beliebige dieser Aufgaben die Dauer $\Delta > 0$ bedient wurde, ist ihre Leistung gefallen, sie muß zugunsten einer anderen abgebrochen werden. Dasselbe Schicksal erleidet kurze Zeit später die unterbrechende Aufgabe und das setzt sich so fort. Eine neue ankommende Anforderung desselben Typs darf unterbrechen und wird solange ununterbrochen (falls nicht eine weitere Aufgabe ankommt) bedient, bis sie mit den schon vorhandenen um infinitesmal kurze Zeitscheiben Δ konkurrieren muß.

Die Bedienung beider Typen von Aufgaben des Beispiels in Bild 7 erfolgt so:

* Aufgaben des Typs i=1 mit $t \leq t_1$ (Klasse 1) werden bedient, als gäbe es nur diesen einen Typ. Falls solche nicht warten, konkurrieren
* Aufgaben des Typs 1 mit $t_1 < t \leq t_2$ und des Typs 2 mit $0 \leq t \leq t_2$ (zusammen: Klasse 2), u.U. mit kleinen Zeitscheiben Δ um Bedienung. Falls weder Aufgaben der Klassen 1 noch 2 warten, konkurrieren
* Aufgaben des Typs 1 mit $t_2 < t \leq t_3$ und des Typs 2 mit $t > t_2$ (zusammen: Klasse 3). Falls die Klassen 1 bis 3 unbesetzt sind, konkurrieren
* Aufgaben des Typs 1 mit $t > t_3$ (Klasse 4) allein untereinander.

Kommt während der Bedienung von Aufgaben der Klasse j eine Aufgabe der Klasse $i \leq j$ an, dann wird die solange bedient, bis sie dieselbe Bediendauer t aufweist (oder fertiggestellt wird) wie die vorher bedienten Aufgaben der Klasse j. Diese Bedienung kann durch weitere Ankünfte unterbrochen werden. Die Bedienung von Aufgaben mit H_r-Verteilung erfordert sehr häufig Unterbrechungen und führt in der praktischen Anwendung zu einem unvertretbar hohen Verwaltungaufwand. Dieses Problem wird entschärft, wenn man -wie allgemein üblich- anstelle sehr kleiner Rechenfristen Δ endliche Zeitscheiben Q zuteilt. Dabei geht die Optimalität der Bedienstrategie verloren. Bei geschickter Wahl der Zeitscheibendauer Q bleibt die Strategie "praktisch kostenoptimal" [7].

Ein anderer Weg, dieses Problem zu meistern, ist durch Ersatz der H_r-Verteilungen durch geeignete stückweise exponentielle (Piecewise Markovian-n) Verteilungen n-ter Ordnung möglich.

4.3 Stückweise exponentielle Verteilungen

PM-n Verteilungen wurden in [3], [4] eingeführt, um zu praktikablen optimalen Bedienstrategien und berechenbaren Wartemodellen zu kommen. Geht man davon aus, daß in Modellen vorkommende H_r-Verteilungen Näherungen für in der praktischen Anwendung vorkommende Verteilungen sind, dann spricht nichts dagegen, andere vergleichbare Funktionen zur Approximation zu wählen, vgl. Bild 5e. Verteilungsfunktion und Verteilung lauten [4]

$$B(t) = 1-\exp[-\sum_{k=1}^{j} t_{g_k}(\mu_k-\mu_{k+1})-\mu_{j+1}t];$$

$$b(t) = \mu_{j+1} \exp[- \sum_{k=1}^{j} t_{gk}(\mu_k - \mu_{k+1}) - \mu_{j+1} t]$$

$$\text{für } t_{gj} < t \leq t_{gj+1}, \quad j=0,1,\ldots n;$$

$$0 = t_{go} < t_{g1} < \ldots t_{gn} < t_{gn+1} = \infty .$$

Für n=0 erhält man die M-Verteilung.
Wir beschränken uns hier auf den Fall der Nachbildung des Verlaufs von H_r-Verteilungen und machen eine geeignete Einschränkung:

$$\mu_1 > \mu_2 > \ldots \mu_i > \ldots > \mu_{n+1} .$$

Unter dieser Voraussetzung fällt für alle PM-n Verteilungen die Leistung mit wachsender Bediendauer t treppenförmig, ist also stückweise konstant. Für eine PM-Verteilung ist

$$\varphi^*(i,t) = \begin{cases} g_i\mu_{i1} & \\ g_i\mu_{i2} & \\ \vdots & \text{für} \\ g_i\mu_{in} & \\ g_i\mu_{i(n+1)} & \end{cases} \quad \begin{array}{l} t+\delta \leq t_{gi1} \\ t+\delta \leq t_{gi2} \\ \vdots \\ t+\delta \leq t_{gin} \\ t > t_{gin},\ \delta \text{ beliebig} \end{array}$$

(Dabei ist δ die Rechenfrist, die zu dieser Lösung führt; andere (größere) Fristen sind nicht kostenminimal).

Da es sich um eine <u>stückweise</u> <u>exponentielle</u> Verteilung handelt, ist zu erwarten, daß die Leistung sich in Abhängigkeit der Bediendauer t so verhält, wie das dem zugehörigen <u>exponentiellen</u> Stück $t_{gj} < t \leq t_{gj+1}$ der Bediendauer-Verteilung entspricht.

Bild 8 zeigt ein Beispiel mit zwei Typen von Aufgaben i=1,2 und verschiedenen PM-1 Verteilungen. Dieses Beispiel entspricht dem von Bild 7.
Aufgaben mit gleicher Leistung sind gleichwertig für die kostenminimale Bearbeitung. Ihre Reihenfolge ist beliebig, z.B. ist die Bedienstrategie FCFS optimal. Wählt man für gleichberechtigte Aufgaben die Reihenfolge FCFS, dann gibt es nur Zeitscheiben der Dauer t_{g1} für Aufgaben des Typs 1 mit t=0 (wird nie unterbrochen), t_{g2} für Aufgaben des Typs 2 mit $t < t_{g2}$ (wird bei Ankunft einer Aufgabe vom Typ 1 unterbrochen, später fortgesetzt) oder q = unbegrenzt (wobei Aufgaben vom Typ 2 bevorzugt vor denen vom Typ 1 fertiggestellt werden).

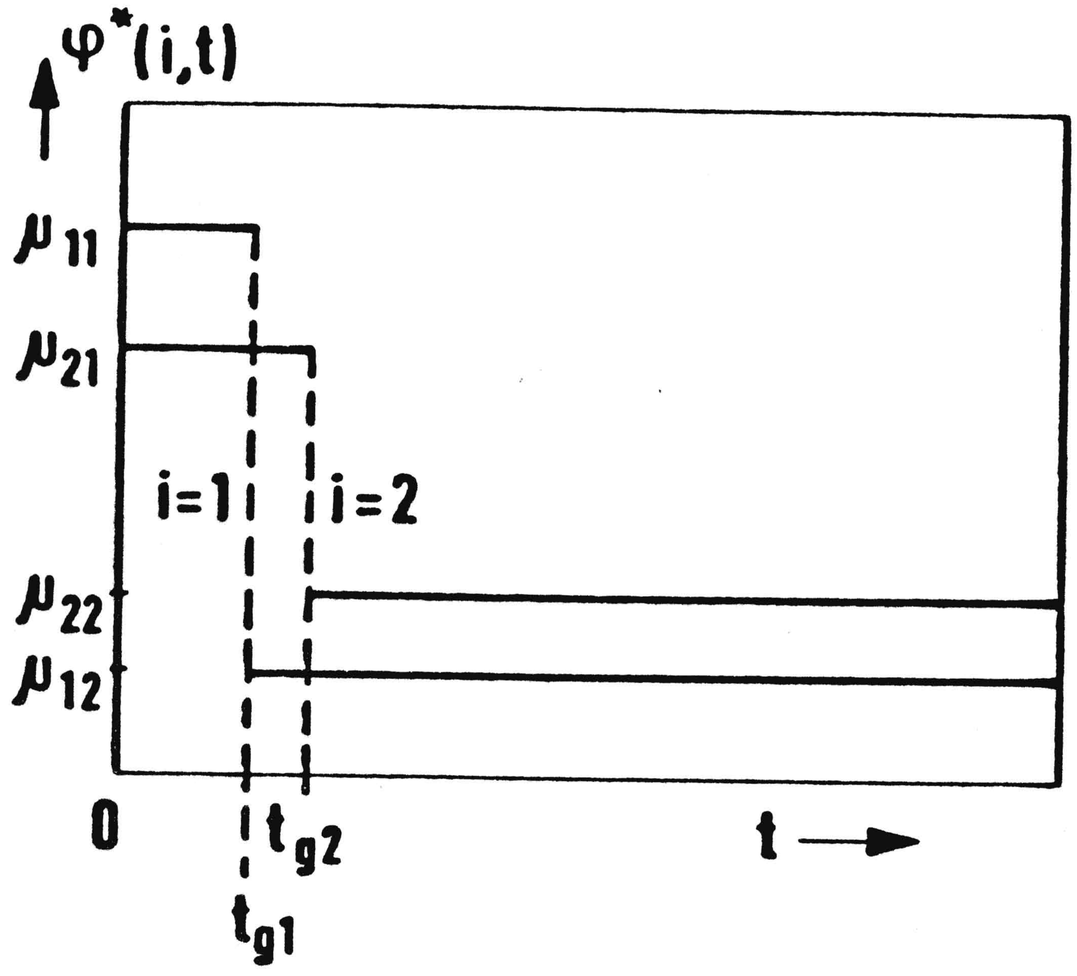


Bild 8: Maximale Leistung für zwei stückweise exponentielle VFn 1. Ordnung über der verbrauchten Prozessorzeit t.

Die unbegrenzte Zeitscheibe q wird bei Ankunft einer Aufgabe unterbrochen.

Anstelle des komplizierten Bedienungsschemas bei H_r-Bediendauerverteilungen mit zeitweise sehr vielen Unterbrechungen und sehr kleinen Zeitscheiben sind bei vergleichbaren PM-1 Verteilungen nur wenige Unterbrechungen nötig.

Man muß sich klar darüber sein, daß die Wahl der Bedienzeiten t_{gk} (vgl. Bild 5e) bei der Konstruktion einer PM-n Verteilung gleichzeitig die optimalen Zeitscheibenlängen $t_{gk+1} - t_{gk}$ festlegt. Wegen der großen möglichen Zahl der Parameter bei PM-n Verteilungen gibt es immer viele gleich gute Lösungen, eine gegebene statistische Verteilung zu approximieren. Das gilt jedoch allgemein auch für die Approximation durch H_r-Verteilungen. Für PM-n Verteilungen sollte eine Näherung mit wenigen exponentiellen Stücken angestrebt werden.

Während man bei Voraussetzung von H_r-Bediendauerverteilungen minimale Zeitscheibendauern Q definieren mußte, um eine praktikable nur näherungsweise optimale Bedienstrategie zu erhalten, führt die Voraussetzung von PM-n Verteilungen auf eine optimale Bedienstrategie mit großen Zeitscheiben und wenigen Unterbrechungen.

5. ZUR ANALYSE VON WARTEMODELLEN MIT OPTIMALER BEDIENSTRATEGIE

Ergebnisse von Überlegungen zu optimalen Bedienstrategien haben besonders dann praktische Bedeutung, wenn

- die resultierenden Strategien anwendbar sind und
- Zahlenwerte für die optimierten charakteristischen Größen berechenbar sind.

Im Rahmen der hier betrachteten Modelle mit unbegrenztem Warteraum und einem Prozessor sind charakteristische Größen des Anwendungsbeispiels nach Abschnitt 4.1 für konstante Kostenfunktionen nur geschlossen berechenbar, wenn keine Unterbrechungen zugelassen sind: Dann ist die Anfangsleistung $\varphi^*(i,o)$ unbearbeiteter Aufgaben des Typs i(i=1,2,...N) alleiniges Entscheidungskriterium für die Bedienstrategie und Aufgaben desselben Typs werden FCFS bedient.

Diese (wegen des Verbots von Unterbrechungen) nicht optimale Strategie führt unmittelbar auf das bekannte Modelll M/G/1 mit statischen nicht-unterbrechenden Prioritäten, dessen Ergebnisse z.B. in [5, Abschnitt 4] zu finden sind.

Es erscheint nicht aussichtslos, auch für die optimale unterbrechende Strategie charakteristische Größen zu bestimmen. Bisher sind jedoch keine Ergebnisse bekannt.

Streuen die Bediendauern stark, wie in den Abschnitten 4.2 und 4.3 angenommen, dann gelingt eine analytische Berechnung unter optimaler Bedienstrategie, wenn man nur einen Typ von Aufgaben mit PM-n Verteilung zuläßt [4]. Der Fall mehrerer konkurrierender Typen von Aufgaben mit unterschiedlichen PM-Verteilungen kann näherungsweise behandelt werden, indem man eine gemeinsame PM-Verteilung ersatzweise bildet, die für alle Aufgaben gilt. Dann ist das Problem auf den Fall mit nur einem Typ von Aufgaben zurückgeführt. Der Näherungscharakter liegt darin, daß man ursprünglich verfügbare Information über die Typ-Zugehörigkeit einer Aufgabe nicht nutzt und dabei z.B. unbearbeiteten Aufgaben eine erste Zeitscheibe $t_{gk} = \min \{ t_{gi}, i=1...N\}$ gleichberechtigt zuteilt (vgl. dagegen die optimale Zuteilung des Beispiels nach Bild 8, Abschnitt 4.3).

In [6] ist gezeigt, daß die Berechnung charakteristischer Größen auch noch gelingt, wenn N Typen von Aufgaben konkurrieren, wobei jeder Typ eine beliebige PM-n Bediendauer-VF haben darf, jeder Typ aber eine eigene statische unterbrechende Priorität besitzt (vgl. Bild 2).

Für den Fall hyperexponentiell verteilter Bediendauern ist kein einziges Modell bekannt, für das die Berechnung charakteristischer Grössen bei optimaler Bedienstrategie gelungen ist.

6. VERALLGEMEINERUNG

Nach [2] ist das Verfahren zur Bestimmung einer optimalen Bedienstrategie über den Leistungsvergleich konkurrierender Aufgaben auch optimal, wenn anstelle von Poisson beliebige homogene Ankunftsprozesse vorliegen. Das Verfahren bleibt optimal, wenn der Ankunftsprozeß nicht homogen aber unabhängig von gegenwärtigen oder vergangenen Systemzuständen ist und die Leistung aller Aufgaben monoton schwach fällt. Im allgemeinen unbeschränkten Fall G/G/1 (ohne Beschränkungen für Ankunftsprozesse und Bediendauerverteilungen) sind nachweislich Bedienstrategien optimal, die Prognosen über zukünftige Ankünfte verarbeiten.

LITERATURZITATE

[1] Olivier, G.: Optimale Zeitzuteilung für wartende Rechenaufgaben. Elektron. Rechenanl. 9 (1967), H.5, S.218-224.

[2] Olivier, G.: Kostenminimale Prioritäten in Wartesystemen vom Typ M/G/1. Elektron. Rechenanl. 14 (1972), H.6, S.262-271.

[3] Marte, G.: Optimal time scheduling for time-shared computer systems with piecewise exponential computing-time distribution. Intern. Computing Symposium, Bonn, Mai 1970, Vorabdruck Bd.1, S. 182-206.

[4] Walke, B., Küspert, H.J.: Teilnehmerrechensysteme: Mittlere Verweilzeiten bei optimaler Rechenzeitzuteilung. Elektron. Rechenanl. 13 (1971), H.5, S.193-199.

[5] Walke, B.: Realzeitrechner-Modelle, Theorie und Anwendung. R. Oldenbourg Verlag München Wien, Reihe Datenverarbeitung, 1978.

[6] Walke, B., Static-priority queue with time slicing. 9th Intern. Teletraffic Congress, Oct. 1979, Torremolinos, Spain.

[7] Walke, B.: Entwicklung optimaler Zuteilungsstrategien für Rechnermodelle durch Simulation und Rechnung. Elektron. Rechenanl. 16 (1974), H.1, S.9-17.