NTG - Fachtagung "Teilnehmer - Rechensysteme"
in Erlangen vom 20. bis 22. September 1967

Wartezeiten und Prioritäten

Werner Wagner

Institut für Nachrichtenvermittlung und Datenverarbeitung
der Universität Stuttgart (Technische Hochschule)

1. Einleitung

Um die Jahrhundertwende wurden automatische Fernsprechvermittlungsämter eingeführt. Im Jahre 1917 veröffentlichte der dänische Mathematiker Agner Krarup ERLANG eine grundlegende Arbeit, in der er Vermittlungssysteme und den Fernsprechverkehr behandelte / 1, 8/. In der sich immer mehr entfaltenden Verkehrstheorie hat von da an die Untersuchung von Warteschlangen einen breiten Raum eingenommen. Warteschlangen treten nicht nur in Fernsprech- und Fernschreibvermittlungsstellen auf. Wir begegnen ihnen an vielen Stellen, wo Dienstleistungen gewünscht werden, z.B. vor Fahrkartenschaltern oder bei Personenaufzügen. Auch in Datenverarbeitungszentren kommt es zu Wartezeiten. Betrachten wir eine Datenverarbeitungsanlage, an die mehrere Konsolen angeschlossen sind. Weil in einem bestimmten Zeitpunkt nur eine Anforderung von einer Baugruppe der Datenverarbeitungsanlage, z.B. dem Rechenwerk, abgefertigt werden kann, müssen die anderen Anforderungen warten. Sind die Wartezeiten für alle Benützer genügend klein, dann ist der Wartesystemcharakter von außen nicht zu spüren.

Innerhalb einer Rechenanlage müssen Anforderungen warten, die eine bestimmte Baugruppe oder ein bestimmtes Werk benötigen. Der nächste Befehl oder der nächste Programmabschnitt erfordert das Rechenwerk oder das Bereitstellen eines Speicherbezirks oder eines Ein- oder Ausgabekanals.

Allgemein soll von Anforderungen und Funktionseinheiten gesprochen werden. Ist die erforderliche Funktionseinheit belegt, dann beginnt die Anforderung eine Wartezeit. Man findet im Schrifttum verschiedene Bezeichnungen für die Funktionseinheiten, z. B. Kanal, Leitung, Abfertigungsplatz oder Bedienungsstation (server).

1.1 Belegungsdauern und Einfallprozeß

Wovon hängen die Wartezeiten ab?

a. Es kommt auf die Belegungsdauern der Anforderungen an, die von der betrachteten Einheit zuerst abgefertigt werden müssen, bevor die betrachtete Anforderung an der Reihe ist. Die mittlere Wartezeit ist umso größer, je größer die mittlere Belegungsdauer ist. Die Wahrscheinlichkeitsverteilung der Belegungsdauern beeinflußt die Wartezeiten. Die Belegungszeiten einer Einheit können konstant sein, oder sie können z.B. negativ exponentiell um den Mittelwert verteilt sein. Für verschiedene Typen von Anforderungen können die Belegungsdauerverteilungen verschieden sein. Die meisten Fälle werden durch eine Verteilung zwischen den ersten beiden Annahmen wirklichkeitstreu beschrieben werden. Man nähert die tatsächlich auftretenden mittleren Wartezeiten mit den Wartezeiten bei diesen beiden Belegungsdauerverteilungen an.

b. Das zeitliche Eintreffen der Anforderungen beeinflußt ebenfalls die Wartezeiten. Wenn die mittlere Zahl der Anforderungen in der Zeiteinheit steigt, wächst die mittlere Wartezeit. Als einfache Annahme setzen wir für das Eintreffen einen Poisson-Prozeß voraus. Die Wahrscheinlichkeit, daß im nächsten Zeitintervall eine Anforderung eintrifft, ist konstant. Die Annahme ist erfüllt, wenn z.B. die Zahl der Konsolen, die zu einer Rechenanlage Zugang haben, sehr viel größer als 1 ist, und wenn die Benutzungsdauer der einzelnen Konsole gegenüber der Gesamtbenutzungsdauer der Rechenanlage hinreichend klein ist. Die Annahme eines Poisson-Prozesses ist dann eine brauchbare Näherung. Wenn die Zahl der Quellen, aus denen die Anforderungen entspringen, sehr klein ist, dann treffen die Anforderungen weniger ungleichmäßig ein als im soeben geschilderten Zufallsangebot 1. Art. Über Messungen der Zeitabstände zwischen Ankünften und über eine Approximation der Verteilung berichteten E.G. COFFMAN und R.C. WOOD /3/.

c. Oft steht nur eine Einheit für alle Benützer zur Verfügung, z.B. das Rechenwerk oder genau der Teilspeicher, in dem die erforderlichen Daten stehen. In anderen Fällen werden sich mehrere Einheiten aushelfen. Sind z.B. 2 gleiche Ausgabegeräte angeschlossen, dann können sie einander ersetzen. Ist ein Speicherbezirk zu reservieren, dann ist es belanglos, welcher unter n gleichartigen Speicherbezirken zugeteilt wird. Für die Zahl der Einheiten sind deshalb die Fälle n=1 und n>1 zu unterscheiden.

1.2 Die gesuchten Größen

Dieselben Probleme treten in der Vermittlungstechnik auf. Rufe kommen an und suchen eine freie Leitung im Bündel von n Leitungen. Können die Rufe nicht durchgeschaltet werden, dann müssen sie warten. Die Verkehrstheorie gibt die Berechnungsverfahren für die interessierenden Größen:

1. Die Wartewahrscheinlichkeit W ist die Wahrscheinlichkeit, daß eine ankommende Anforderung warten muß.
2. Die mittlere Wartezeit τ_W ist der Erwartungswert der zufälligen Wartezeiten der Anforderungen, die überhaupt warten müssen. Die mittlere Belegungsdauer ist die Zeiteinheit.
3. Die Mindestwertverteilungsfunktion $W(>\tau)$ gibt die Wahrscheinlichkeit, daß eine wartende Anforderung länger als die Zeit τ warten muß.

Wartewahrscheinlichkeit und mittlere Wartezeit lassen sich leichter berechnen als die Verteilungsfunktion der Wartezeiten. Die Aufgabe wird aber oft so gestellt, daß eine bestimmte Höchstwartezeit nur mit einer vorgeschrieben geringen Wahrscheinlichkeit überschritten werden darf.

Manche Systeme begrenzen die Wartezeit. Die Zahl der Warteplätze kann beschränkt sein. Aber dann werden Anforderungen abgewiesen, sie erhalten wie im Vermittlungssystem Besetztzeichen und werden nicht abgefertigt.

2. Prioritäten und Abfertigungsreihenfolge

Oft ist es nicht angebracht, alle Anforderungen als gleichwertig zu betrachten. Durch Warten auf eine Einheit E_2 kann eine besonders wichtige Einheit E_1 länger belegt sein, weil sie mit der Einheit E_2 zusammenarbeiten soll. Man möchte diese Art von Anforderungen bevorzugen, um kurze Wartezeiten zu erreichen. Im on-line-Betrieb sind manche Anforderungen sehr eilig, weil sie z.B. durch Störungen der zu überwachenden Einrichtungen ausgelöst wurden. Andere Anforderungen sind weniger dringend, weil sie z.B. routinemäßigen Überprüfungen dienen.* Darum möchte man gewissen Anforderungen einen Vorrang vor anderen Anforderungen zuerkennen und führt Prioritäten ein. Ohne Prioritäten muß das System so ausgelegt werden, daß für alle Anforderungen die Wartezeiten nur mit genügend kleiner Wahrscheinlichkeit eine Zeitschranke überschreiten. Durch Prioritäten wird das System an die speziellen Anforderungen angepaßt.

Welche unter den wartenden Anforderungen kommt zuerst daran, eine soeben frei gewordene Einheit zu belegen? Zunächst wird auf die Priorität geachtet. Die Abfertigungsreihenfolge beeinflußt die Verteilungsfunktion der Wartezeiten. Gebräuchlich ist die Annahme, daß die wartenden Anforderungen in der Reihenfolge ihrer Ankunft abgefertigt werden. Das gilt für die wartenden Anforderungen jeder einzelnen Prioritätsklasse.

Die Zahl der Probleme ist sehr groß: man kann verschiedene Belegungsdauerverteilungen, Einfallprozesse, Zahl von Einheiten, Prioritäteneinteilungen und Abfertigungsreihenfolgen kombinieren.

*Es können auch den verschiedenen Konsolen von vornherein verschiedene Prioritäten zugeteilt sein.

3. Wartesystem mit negativ exponentieller Verteilung der Belegungsdauern

Wir gehen von folgendem Wartesystem aus: Jede Anforderung benötigt irgendeine aus insgesamt n gleichwertigen Funktionseinheiten. Die vollkommene Aushilfe heißt in der Verkehrstheorie vollkommene Erreichbarkeit. Die Wahrscheinlichkeit, daß die Belegung einer Einheit in einem beliebigen infinitesimal kleinen Zeitabschnitt endet, sei konstant und unabhängig davon, wie lange die Belegung bereits andauert und welche Priorität sie besitzt; für alle Belegungen gilt die gleiche negativ exponentielle Verteilung. Die mittlere Belegungsdauer wählen wir als Zeiteinheit.

Die Anforderungen werden in die Prioritätsklassen 1, 2, ... k, ... K eingestuft. Anforderungen höchster Dringlichkeit gehören zur 1. Prioritätsklasse; mit abnehmender Dringlichkeit wächst der Prioritätsklassenindex k. Während der Zeiteinheit, der mittleren Belegungsdauer, treffen im Mittel A_k Anforderungen der Prioritätsklasse k ein. A_k ist das Angebot in der Prioritätsklasse k. Die Wahrscheinlichkeit, daß in einem beliebigen infinitesimal kleinen Zeitabschnitt eine Anforderung der Prioritätsklasse k ankommt, sei konstant, unabhängig von der Vorgeschichte; für jede Priorität wird ein Poisson-Prozeß oder Zufallsangebot 1. Art angenommen.

Anforderungen höherer Priorität reihen sich in der Warteschlange vor Anforderungen von geringerer Priorität ein. Anforderungen, die bereits eine Einheit belegt haben, werden auch dann nicht unterbrochen, wenn neu ankommende Anforderungen höherer Priorität warten müssen. *

3.1 Zustandswahrscheinlichkeiten

Es ist eine grundlegende Methode der Verkehrstheorie, die Gleichungen für die Zustandswahrscheinlichkeiten aufzustellen. Eine ankommende Anforderung der Prioritätsklasse k (in Bild 1 z.B. k=2) stellt sich zu den ζ_k (im Beispiel $\zeta_2=4$) bereits wartenden Anforderungen aus den Prioritätsklassen 1 bis k. Es ist $p(\zeta_k \leq k)$ die Wahrscheinlichkeit dafür, daß ζ_k Anforderun-

* Jede Anforderung wartet so lange, bis sie eine Funktionseinheit belegen darf; auch bei langen Wartezeiten kommt es nie zu Verzichten.

gen aus den Prioritätsklassen $\leq k$ warten und daß zugleich alle n Einheiten belegt sind.

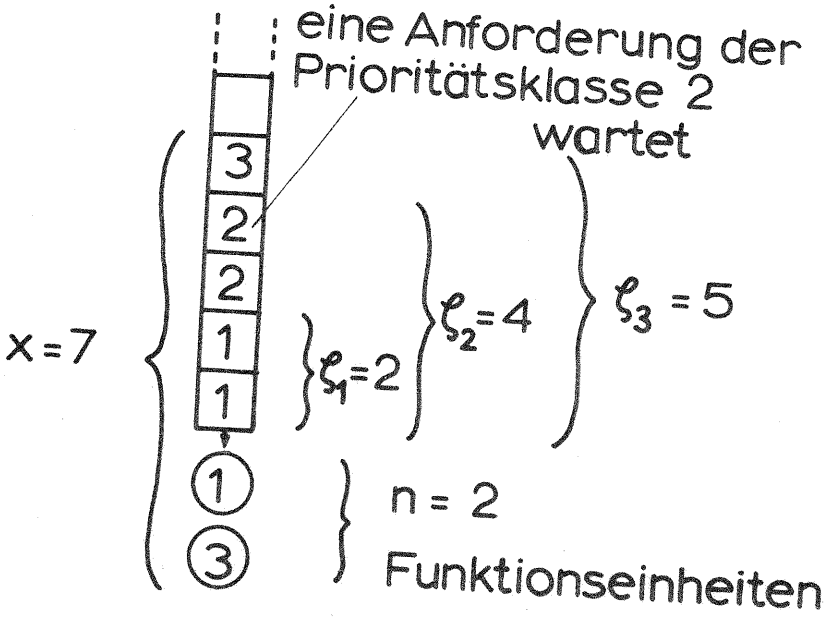

Bild 1. Beispiel für einen Augenblickszustand

Die Zahl der Wartenden aus den Prioritätsklassen $\leq k$ wird durch eine Anforderung aus diesen Prioritätsklassen $\leq k$ um 1 erhöht und andererseits um 1 vermindert, wenn eine der n bestehenden Belegungen endet und die vorderste der ϱ_k wartenden Anforderungen eine Einheit belegen kann.

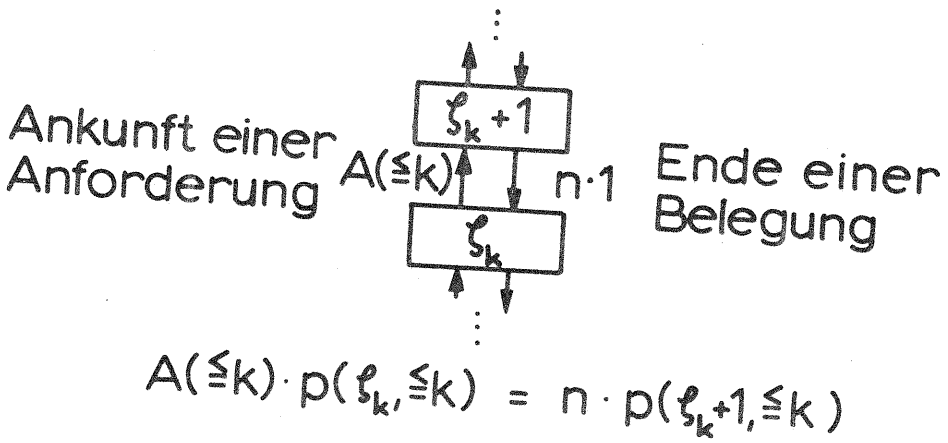

Bild 2. Übergänge zwischen benachbarten Zuständen

Wir betrachten den stationären Fall, daß die Zustandswahrscheinlichkeiten $p(\zeta_k, \leqq k)$ vom Beobachtungszeitpunkt unabhängig sind. Mit der Rufwahrscheinlichkeitsdichte oder, mit anderen Worten, mit dem

$$A(\leqq k) = A_1 + A_2 + \ldots + A_k \qquad (1)$$

Angebot

und der Endewahrscheinlichkeitsdichte $n \cdot 1$ von n unabhängigen Belegungen ergibt sich

$$A(\leqq k) \cdot p(\zeta_k, \leqq k) = n \cdot p(\zeta_k + 1, \leqq k) \qquad \text{für } \zeta_k = 0, 1, 2, \ldots \qquad (2)$$

Ein solches System von Zustandsgleichungen bezeichnete A.K. ERLANG als "Bedingung für die Erhaltung des statistischen Gleichgewichts". (Zur ausführlicheren Begründung des Ansatzes vgl. z.B. /13/)

$p(0 \leqq k)$ ist die Wahrscheinlichkeit, daß keine Anforderung aus den Prioritätsklassen $\leqq k$ wartet und daß alle n Einheiten belegt sind. Deshalb gilt die Gleichung für $\zeta_k = 0$ ebenfalls. Sie soll für unbegrenzt große Werte ζ_k gelten. Ein unbegrenzt großer Wartespeicher wird in der Technik nahezu erreicht. Zum Beispiel können auf Magnetband sehr große Folgen von Daten gespeichert werden, die auf den Transfer in die zugewiesenen Kernspeicherbezirke warten.

Der Zustand des Wartesystems in einem bestimmten Zeitpunkt ist weiterhin gegeben durch die Gesamtzahl x von wartenden Anforderungen und Belegungen in den Einheiten (im Beispiel x= 7 , vgl. Bild 1). Die Wahrscheinlichkeit ist p_x. In den Gleichungen für die Wahrscheinlichkeiten der Zustände, daß ζ_k Anforderungen der Prioritätsklassen $\leqq k$ vor den belegten Einheiten warten, ist die Bedingung enthalten, daß in allen Fällen n Einheiten belegt sind.

$$\sum_{\zeta_k=0}^{\infty} p(\zeta_k, \leqq k) = \sum_{x=n}^{\infty} p_x = E(n, A) = \begin{cases} E_{2,n}(A) & \text{für } A = A(\leqq K) < n \\ & \text{2. Formel von ERLANG /1,8,13/} \\ 1 & \text{für } A = A(\leqq K) \geqq n \end{cases} \qquad (3)$$

Für die Zufallsvariable x von Anforderungen im System, d.h. für den momentanen Gesamtverkehr, ergibt sich nur für $A<n$ ein stationärer Prozeß mit konstanten, von der Zeit nicht mehr abhängigen Wahrscheinlichkeiten und Erwartungswerten. Für $A \geqq n$ wächst der Erwartungswert der Warteschlangenlänge dauernd an, das System ist gesättigt. /11/

Mit dem auf eine Funktionseinheit bezogenen Angebot der Prioritätsklassen $\leq k$

$$\alpha(\leq k) = \frac{A(\leq k)}{n} \qquad (4)$$

lautet die Formel für die gesuchte Wahrscheinlichkeit

$$p(\zeta_k \leq k) = [\alpha(\leq k)]^{\zeta_k} \cdot [\,1-\alpha(\leq k)\,] \cdot E(n,A) \text{ für } A(\leq k) < n \qquad (5)$$

Für n=1 gaben S.A. DRESSIN und E. REICH die Beziehung an. /7,13/ Für die Prioritätsklassen 1, 2, ... k stellt sich ein stationärer Zustand ein, wenn $A(\leq k) < n$ ist, auch dann, wenn insgesamt kein stationärer Zustand mehr erreicht wird, weil $A \geq n$ ist. Die Zahl der wartenden Anforderungen und der Belegungen in den Einheiten, beide aus den Prioritätsklassen $\leq k$, schwankt dann um einen zeitlich konstanten Mittelwert, dagegen nimmt die Zahl der wartenden Anforderungen der letzten Prioritätsklassen immer mehr zu.

3.2 Wartewahrscheinlichkeit und mittlere Wartezeit

Die Wartewahrscheinlichkeit W ist unabhängig von der Priorität einer neu ankommenden Anforderung, weil bestehende Belegungen in dem betrachteten Modell in keinem Fall unterbrochen werden.

$$W = E \qquad \text{für } k = 1, 2, \ldots K \qquad (6)$$

Während der mittleren Wartezeit τ^*_{wk}, bezogen auf die mittlere Belegungsdauer als Zeiteinheit und auf alle Anforderungen der Prioritätsklasse k, treffen im Mittel $\tau^*_{wk} \cdot A_k$ neue Anforderungen der Prioritätsklasse k ein. Diese Anforderungen erhalten im Mittel bei stationärem Prozeß gerade die mittlere Warteschlangenlänge Ω_k aufrecht.

$$\Omega_k = \sum_{\zeta_k=0}^{\infty} \zeta_k \cdot p(\zeta_k \leq k) - \sum_{\zeta_{k-1}=0}^{\infty} \zeta_{k-1} \cdot p(\zeta_{k-1} \leq k-1) = \tau^*_{wk} \cdot A_k \qquad (7)$$

Die mittlere Wartezeit τ_{wk}, nur auf die wartenden Anforderungen der Prioritätsklasse k bezogen, ist mit $\alpha(\leq k-1) = \alpha(<k)$ aus (5) und (7)

$$\tau_{wk} = \frac{\tau^*_{wk}}{W} = \frac{1}{n \cdot [1-\alpha(\leq k)] \cdot [1-\alpha(<k)]} \text{ für } \alpha(\leq k) < 1 \qquad (8)$$

Für Anforderungen höchster Priorität ist k=1, $\alpha(\leq 1) = \alpha_1$, $\alpha(<1) = 0$,

$$\tau_{w1} = \frac{1}{n \cdot (1-\alpha_1)} \qquad (9)$$

Gibt es nur diese eine Klasse von Anforderungen, dann erhält man mit $\alpha_1=\alpha$ ERLANGS bekannte Formel für die mittlere Wartezeit der Wartenden /1,8,13/

$$\tau_w = \frac{1}{n\cdot(1-\alpha)} = \frac{1}{n-A} \qquad (10)$$

3.3 Grenzwerte der mittleren Wartezeiten

Die geforderten Werte der mittleren Wartezeiten lassen sich unter Umständen dadurch erfüllen, daß man den ankommenden Strom der Anforderungen auf andere Weise in Prioritätsklassen einteilt. Wir fragen nach den Grenzwerten. Die kleinstmögliche Wartezeit für Anforderungen höchster Priorität ist gegeben, wenn nie mehr als eine Anforderung höchster Priorität wartet. Sie steht an der Spitze der Warteschlange und wartet bis zum nächsten Belegungsende. Damit gilt als untere Grenze:

$$\lim_{A_1\to 0} \tau_{w1} = \tau_{wu} = \frac{1}{n} \qquad (11)$$

Für die obere Grenze der mittleren Wartezeiten betrachten wir solche Anforderungen, die sich, wenn sie warten müssen, an das Ende der Warteschlange stellen. Jedesmal, wenn eine Funktionseinheit frei wird, rücken sie um einen Warteplatz vor. Sie lassen aber alle während ihrer Wartezeit eintreffenden Anforderungen vor sich in die Warteschlange hinein. Eine solche Anforderung belegt nur dann eine frei werdende Funktionseinheit, wenn außer ihr überhaupt keine andere Anforderung wartet. Für die obere Grenze ergibt sich (vgl. E.VAULOT /13/) mit $\alpha=\alpha(\leq K)$:

$$\lim_{A_K\to 0} \tau_{wk} = \tau_{wo} = \frac{1}{n}\cdot\frac{1}{(1-\alpha)^2} \qquad (12)$$

3.4 Ein Beispiel

In einem Beispiel ist die Prioritäteneinteilung

Prioritätsklasse k		Anteil der Anforderungen $\frac{A_k}{A}$
1	"sofort"	3 %
2	"dringend"	25 %
3	"Routine"	32 %
4	"Zurückstellung"	40 %

Bei größer werdendem Angebot zeigt sich der Einfluß des Einteilens in Prioritätsklassen immer deutlicher. Wenn die mittlere Wartezeit τ_{wk} aufgezeichnet wird (Bild 3), ergibt sich, daß für die höchste Priorität k=1 auch bei großen Angeboten bis zu A≈n die Mindestwartezeit (Erwartungswert) $\frac{1}{n}$ höchstens um 3,1% überschritten wird. Die mittleren Wartezeiten werden durch Einführen von Prioritätsklassen für die höchsten Prioritäten deutlich verringert. Das System verharrt für die sehr wichtigen Anforderungen auch bei höheren Angeboten im stationären Zustand. Die Überlastungsunempfindlichkeit für die hochwertigen Anforderungen ist erwünscht. Die Wartezeitverminderung gilt für alle Einheitenzahlen n in gleichem Maße, wenn die Belegungsdauern der negativ exponentiellen Verteilung gehorchen.

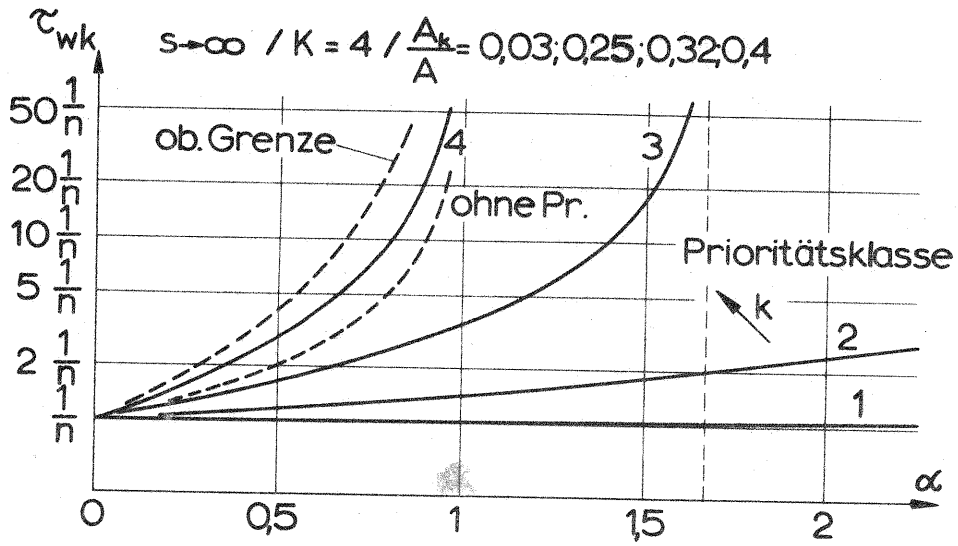

Bild 3. Mittlere Wartezeit τ_{wk} der wartenden Anforderungen der Prioritätsklasse k, abhängig von dem auf eine Funktionseinheit bezogenen Gesamtangebot α

4. Beliebige Verteilung der Belegungsdauern

4.1 Eine Funktionseinheit

Um mit dem Wartesystem ohne Prioritäten zu vergleichen, bilden wir aus den Gleichungen (8) und (10) das Verhältnis

$$\frac{\tau_{wk}}{\tau_w} = \frac{1-\alpha}{[1-\alpha(\leq k)]\cdot[1-\alpha(<k)]} \qquad (13)$$

Die Formel (13) für die relative Verlängerung der mittleren Wartezeit τ_{wk} durch Einteilen in Prioritätsklassen gilt

1. für das Wartesystem mit Poisson-Prozessen für die Ankünfte, mit negativ exponentieller Verteilung der Belegungsdauern und mit $n \geq 1$ Funktionseinheiten.

Sie wurde auf ganz anderem Wege von A. COBHAM hergeleitet /2/.
Die Formel (13) gilt außerdem

2. für das Wartesystem mit Poisson-Prozessen für die Ankünfte, mit beliebiger Verteilung der Belegungsdauern und mit nur 1 Funktionseinheit. /2,11,14,18/

Die Wartewahrscheinlichkeit W ist bei nur einer Funktionseinheit unabhängig von der Verteilung der Belegungsdauern stets gleich dem Angebot A, denn das ist zugleich die Wahrscheinlichkeit, die eine Funktionseinheit belegt anzutreffen.

$$W = A \qquad \text{für } n=1 \qquad (14)$$

Die mittlere Wartezeit τ_w^* aller Anforderungen und die mittlere Wartezeit τ_w der Wartenden ist in dem 2. Wartesystem ohne Prioritäten durch die POLLACZEK-KHINTCHINE-Formel gegeben /13,18/:

$$\tau_w = \frac{\tau_w^*}{A} = \frac{(1+\sigma^2)}{2\cdot(1-A)} \qquad \text{für } n=1 \text{ und } \alpha=A \qquad (15)$$

σ^2 ist die Varianz der Belegungsdauern. Konstante Belegungsdauern bedeuten $\sigma^2=0$, negativ exponentielle Verteilung führt auf $\sigma^2=1$.

4.2 Ein Beispiel für die relative Verlängerung der mittleren Wartezeiten

Wir betrachten 4 Prioritätsklassen; der Anteil $\frac{A_k}{A}$ in den einzelnen Prioritätsklassen ist jeweils 25 %. Verglichen mit der mittleren Wartezeit τ_w ohne Prioritäteneinteilung werden die mittleren Wartezeiten für die ersten beiden Prioritätsklassen, für die Hälfte der Anforderungen, deutlich verringert.

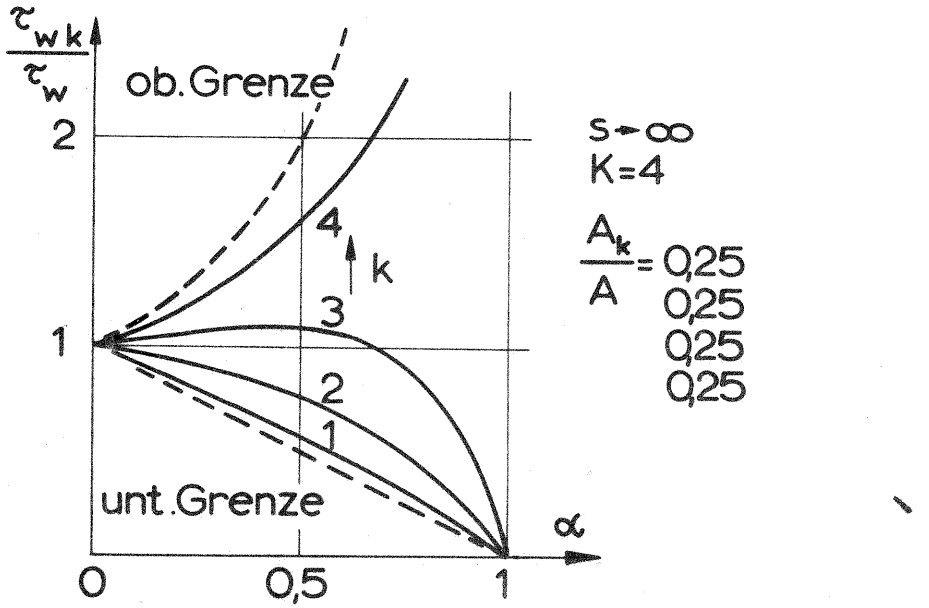

Bild 4. Die relative Verlängerung der mittleren Wartezeiten τ_{wk}/τ_w der wartenden Anforderungen der Prioritätsklasse k hängt von dem auf eine Funktionseinheit bezogenen Gesamtangebot α ab.

4.3 Mehrere Funktionseinheiten

Für mehrere Funktionseinheiten und beliebige Verteilung der Belegungsdauern ist für ein Angebot ohne Prioritäteneinteilung kein einfaches Berechnungsverfahren bekannt; für Systeme mit Prioritäten steht die Lösung aus. Wir haben daher den zeitlichen Verlauf im Wartesystem mit künstlichem Verkehr nach dem zeittreuen Simulationsverfahren /16/ nachgebildet. Wie im vorhergehenden Beispiel betrachten wir 4 Prioritätsklassen; im Mittel ist der Anteil an den ankommenden Anforderungen jeweils

25 %. Die Anforderungen werden von n=2 Funktionseinheiten abgefertigt. Die Testergebnisse sind mit Vertrauensintervallen angegeben, die mit der Studentschen t-Verteilung für die statistische Aussagesicherheit 95 % berechnet wurden /20/. Für die Belegungsdauern wurden 5 verschiedene Verteilungen angenommen.*

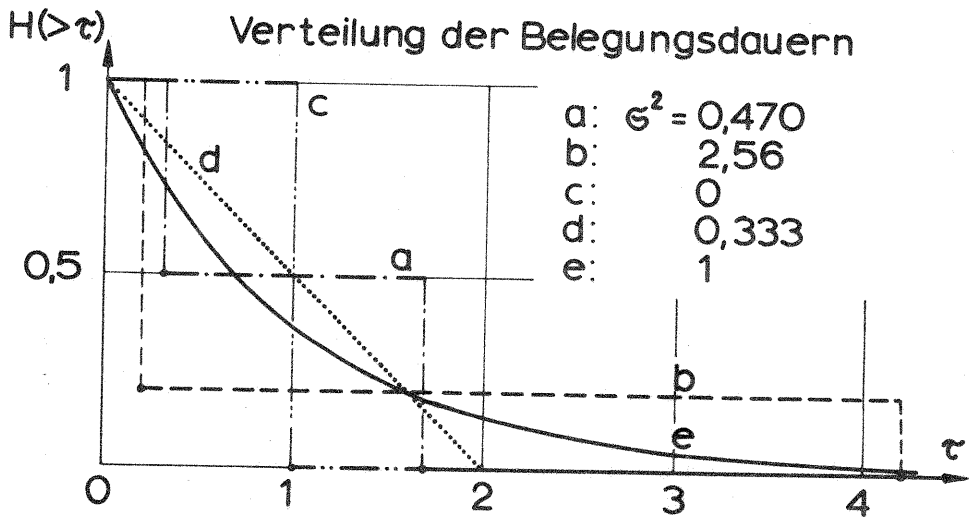

Bild 5. Belegungsdauerverteilungen:

- a Belegungsdauer h_i mit Wahrscheinlichkeit p_i
  $h_1=0,31445$, $p_1=0,5$; $h_2=1,68555$, $p_2=0,5$
- b $h_1=0,2$, $p_1=0,8$; $h_2=4,2$, $p_2=0,2$
- c konstante Belegungsdauer h=1
- d Gleichverteilung der Belegungsdauern zwischen den Grenzwerten 0 und 2
- e negativ exponentielle Verteilung

*Die Verteilung der Belegungsdauern gilt jeweils einheitlich für alle Prioritätsklassen.

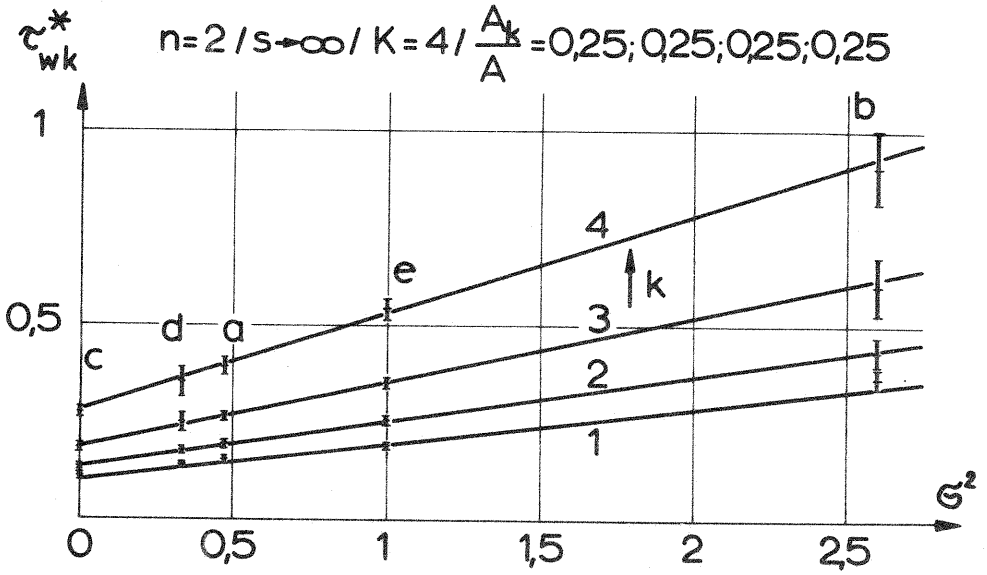

Bild 6. Die mittlere Wartezeit τ^*_{wk} aller Anforderungen der Prioritätsklasse k hängt von der Varianz σ^2 der Belegungsdauern ab. Das Angebot ist A=1,0, die Angebotsanteile der 4 Prioritätsklassen jeweils 25 %. Zahl der Einheiten n=2. Belegungsdauerverteilungen siehe Bild 5.

Wird die mittlere Wartezeit τ^*_w bei konstanter Belegungsdauer mit τ^*_{wc}, bei negativ exponentieller Verteilung der Belegungsdauern mit τ^*_{we} bezeichnet, dann läßt sich Gleichung (15) umformen; für n=1:

$$\tau^*_w = (1-\sigma^2)\cdot\tau^*_{wc} + \sigma^2\cdot\tau^*_{we} \qquad (16)$$

Diese lineare Abhängigkeit wurde von M. BJÖRKLUND und A. ELLDIN heuristisch auf Wartesysteme ohne Prioritäten mit n>1 Funktionseinheiten ausgedehnt /19/. Für beliebige Werte n können τ^*_{wc} und τ^*_{we} exakt berechnet werden /1,13/. Die Hypothese soll jetzt auch für Wartesysteme mit Prioritäten angewendet werden. Weil die Wartewahrscheinlichkeit W nicht von der Prioritätsklasse k abhängt, folgt aus (13) mit $\tau^*_{wk} = W\cdot\tau_{wk}$ und mit (16):

$$\tau^*_{wk} = \frac{1-\alpha}{[1-\alpha(\leq k)]\cdot[1-\alpha(<k)]}\cdot\left[(1-\sigma^2)\cdot\tau^*_{wc} + \sigma^2\tau^*_{we}\right] \qquad (17)$$

wo τ^*_{wc} und τ^*_{we} Funktionen von n und A sind.

Die Hypothese (17) der linearen Abhängigkeit der mittleren Wartezeit τ^*_{wk} aller Anforderungen der Prioritätsklasse k von der Varianz σ^2 der Belegungsdauern wird auf Grund der Tests nicht verworfen (Bild 6). Bei größerer Zahl von Funktionseinheiten, z.B. n=10, weichen die Testergebnisse von der Näherung (17) ab. Es sind weitere Untersuchungen notwendig, die andere Parameter, z.B. die höheren Momente der Verteilung der Belegungsdauern, berücksichtigen.

5. Wartezeitverteilung

Wir kehren zum 1. Wartesystem zurück: n Einheiten, Poisson-Prozesse der Ankünfte und negativ exponentielle Verteilung der Belegungsdauern. Die Wartezeitverteilung $W_k(>\tau)$ ist die gesuchte Wahrscheinlichkeit, daß eine Anforderung der Prioritätsklasse k, die warten muß, mindestens die Zeit τ warten muß.

5.1. Anforderungen höchster Priorität

Die Wartezeitverteilung $W_1(>\tau)$ der wartenden Anforderungen der ersten Prioritätsklasse ist so gegeben, als käme nur das Angebot A_1 aus dieser Prioritätsklasse /14,18/. Während der Wartezeiten müssen alle Anforderungen anderer Prioritäten hinter den Anforderungen der 1. Prioritätsklasse bleiben und beeinflussen die Wartezeiten der wartenden Anforderungen der 1. Prioritätsklasse nicht.

$$W_1(>\tau) = e^{-(1-\alpha_1)\cdot n\cdot\tau} \qquad \text{für } \tau \geq 0 \text{ und } \alpha_1 = \frac{A_1}{n} < 1 \quad (18)$$

5.2 Anforderungen, die nicht zur ersten Prioritätsklasse gehören

Um die Wartezeitverteilung $W_k(>\tau)$ für die anderen Prioritätsklassen $k \geq 2$ zu bestimmen, muß der Warteplatz berücksichtigt werden, auf dem eine Anforderung wartet. Aus den möglichen Übergängen zu benachbarten Warteplätzen innerhalb eines kleinen Zeitabschnitts ergeben sich Beziehungen zwischen den Wahrscheinlichkeiten, von bestimmten Warteplätzen aus mindestens eine ge-

gebene Zeit τ lang warten zu müssen. Mit dem Übergang zu unbegrenzt verkleinertem Zeitabschnitt ergibt sich ein System von Differentialgleichungen. S.A. DRESSIN und E. REICH gelangten zur Laplace-Transformierten der Wahrscheinlichkeitsdichte $w_k(\tau)$, welche die Ableitung der Wartezeitverteilung $W_k(>\tau)$ ist /7,13/. Sie wird hier für $n \geq 1$ umgerechnet angegeben:

$$\omega_k(\sigma) = \int_{\tau=0}^{\infty} e^{-\sigma \cdot n \cdot \tau} \cdot w_k(\tau) \cdot d\tau$$

$$= \frac{2 \cdot (1-\lambda)}{1+\lambda-2 \cdot \mu+\sigma+\sqrt{(1+\lambda+\sigma)^2-4 \cdot \lambda}} \qquad (19)$$

für $\lambda = \alpha(<k) = \alpha(\leq k-1)$; $\quad \mu = \alpha(\leq k) < 1$ (20)

Beim Rücktransformieren von $\omega_k(\sigma)$ ergibt sich für die gesuchte Wartezeitverteilung $W_k(>\tau)$ eine Formel, die umfangreiche Berechnungen erfordert /6/.

Eine Wahrscheinlichkeitsverteilung ist auch durch ihre Momente bestimmt. Man ermittelt deshalb die Anfangsmomente oder gewöhnlichen Momente ν-ter Ordnung $M_{\nu(k)}$ /20/

$$M_{\nu(k)} = \int_{\tau=0}^{\infty} \tau^{\nu} \cdot w_k(\tau) \cdot d\tau \qquad (21)$$

und erhält

$$M_{1(k)} = \tau_{wk} = \frac{1}{n} \cdot \frac{1}{(1-\mu) \cdot (1-\lambda)} \qquad \text{vgl. (8)}$$

$$M_{2(k)} = \frac{2}{n^2} \cdot \frac{1-\lambda\mu}{(1-\mu)^2 \cdot (1-\lambda)^3} \qquad \text{/11,13/} \quad (22)$$

$$M_{3(k)} = \frac{6}{n^3} \cdot \frac{1+\lambda-4\lambda\mu+\lambda\mu^2+\lambda^2\mu^2}{(1-\mu)^3 \cdot (1-\lambda)^5} \qquad (23)$$

Daraus werden die zentralen Momente $\mu_{\nu(k)}$ errechnet. /20/

$$\mu_{\nu(k)} = \int_{\tau=0}^{\infty} (\tau-\tau_{wk})^{\nu} \cdot w_k(\tau) \cdot d\tau \qquad (24)$$

Die Varianz $\mu_{2(k)}$ ist

$$\frac{\mu_{2(k)}}{\tau_{wk}^2} = 1 + 2 \cdot \frac{1-\mu}{1-\lambda} > 1 \qquad \text{für } \mu<1 \text{ und } k=2,3,\ldots K \tag{25}$$

Das 3. zentrale Moment $\mu_{3(k)}$ ergibt sich zu

$$\frac{\mu_{3(k)}}{2\cdot\tau_{wk}^3} = 1 + 3\cdot\lambda\cdot(1-\mu)\cdot\frac{2\cdot(1-\mu)+\mu\cdot(1-\lambda)}{\mu\cdot(1-\lambda)^2} > 1 \qquad \text{für } \mu<1 \text{ und } k=2,3,\ldots K \tag{26}$$

Das Einteilen in Prioritätsklassen vergrößert immer die Varianz und das dritte zentrale Moment der Wahrscheinlichkeitsverteilung der Wartezeiten der wartenden Anforderungen, die nicht die höchste Priorität k=1 besitzen, gegenüber der negativ exponentiellen Verteilung der Wartezeiten, die für das Wartesystem ohne Prioritäten gilt; vgl. Gleichung (18).

5.3 Approximation der Wartezeitverteilung

Als einfache Näherung für Mindestwertverteilungsfunktionen benützt man nach J. RIORDAN /13/ die Summe von 2 Exponentialfunktionen.

$$W_k(>\tau) \approx C\cdot e^{-\frac{\tau}{\tau_{wk}+D_1}} + (1-C)\cdot e^{-\frac{\tau}{\tau_{wk}+D_2}} \tag{27}$$

Die Konstanten C, D_1 und D_2 werden so bestimmt, daß die Anfangsmomente 1. bis 3. Ordnung der exakten Verteilungsfunktion und der Approximation übereinstimmen. Die Funktion (27) ist mit $0<C\leq 1$ als Näherung für Verteilungsfunktionen geeignet, deren Varianz und 3. zentrales Moment größer als bei einer einzelnen negativ exponentiellen Verteilung sind. Mit

$$S = \frac{\frac{1}{6}\cdot M_{3(k)} - 3\cdot\frac{1}{2}\cdot M_{2(k)}\cdot M_{1(k)} + 2\cdot M_{1(k)}^3}{\frac{1}{2} M_{2(k)} - M_{1(k)}^3} \tag{28}$$

bestimmt man die Konstanten C, D_1 und D_2 aus

$$D_{1,2} = S \pm \sqrt{\left(\frac{S}{2}\right)^2 - M_{1(k)}^2 + \frac{1}{2} M_{2(k)}} \tag{29}$$

$$C = \frac{-D_2}{D_1-D_2} \tag{30}$$

Falls $\frac{1}{2} \cdot M_{2(k)} = M_{1(k)}^2$ ist, ergibt sich nur eine Exponentialfunktion

$$W_k(>\tau) \approx e^{-\frac{\tau}{M_{1(k)}}} \qquad (31)$$

Vergleicht man das 4. und 5. Anfangsmoment und den Anfangswert der Wahrscheinlichkeitsdichte $w_k(0)$ der Approximation (27,29,30) mit den entsprechenden aus (19) errechneten exakten Werten, dann findet man, daß die relativen Abweichungen höchstens Werte nahe bei 1 % erreichen, falls $\alpha(<k) \leqq 0,1$ und $\alpha(\leqq k) < 1$ ist; für relative Abweichungen unter 10 % ist der Bereich $\alpha(<k) \leqq 0,3$ und $\alpha(\leqq k) < 1$ erlaubt.* des 4. und 5. Anfangsmoments und des Anfangswerts der Wahrscheinlichkeitsdichte

5.4 Ein Beispiel

Die Anteile der 4 Prioritätsklassen seien gleich wie beim Beispiel im Abschnitt 3.4. Die Diagramme zeigen die Wartezeitverteilungen nach Gleichung (27). Außerdem wurden durch Simulation ermittelte Punkte aus zwei Testläufen eingezeichnet.

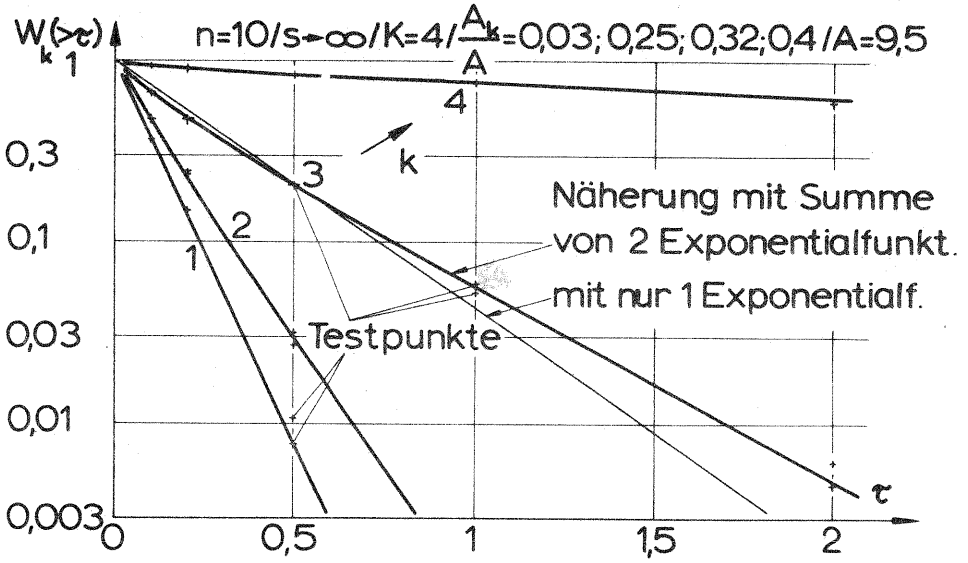

Bild 7. Wartezeitverteilungen $W_k(>\tau)$ für die wartenden Anforderungen der Prioritätsklasse k. Angebot A=9,5 bei n=10 Funktionseinheiten.

*Damit werden vor allem die Werte der Wartezeitverteilung $W_k(>\tau)$ für große Wartedauern τ geprüft.

6. Verschiedene Prioritätenwirksamkeit

6.1 Unterbrechende Priorität

Betont man die Wichtigkeit der Anforderungen der 1. Prioritätsklasse noch mehr, sodaß sie bestehende Belegungen in der Einheit sogar unterbrechen dürfen /5/, dann vermindert man die Wartewahrscheinlichkeit. Nur noch das Angebot A_1 verursacht Belegungen, die zum Warten für die 1. Prioritätsklasse führen können. Die Wartewahrscheinlichkeit ist $E_{2,n}(A_1)<E_{2,n}(A)$. Aber die mittlere Wartezeit der Wartenden und sogar die Wartezeitverteilung der Wartenden der 1. Prioritätsklasse bleiben unverändert.

6.2 Zuteilen der Priorität durch das Rechensystem

Daß die ankommenden Anforderungen bereits einer Prioritätsklasse zugeteilt sind, war bisher vorausgesetzt worden. Andere Verfahren gehen von den einzelnen Belegungsdauern aus. Kurze Belegungen erhalten Vorrang vor langen Belegungen. Jede Anforderung wird in einem Zeitsegment teilweise abgefertigt, dann muß sie erneut auf ein nächstes Zeitsegment warten; dabei kann ihre Priorität verringert werden. /4/

6.3 Endlicher Wartespeicher

Wenn die mittleren Wartezeiten in einem gewissen Bereiche bleiben sollen, dann kann die Zahl der Anforderungen, die in das Rechensystem eingelassen werden, beschränkt werden. Die Priorität beeinflußt dann die Wahrscheinlichkeit $1-B_k$, daß eine Anforderung abgefertigt wird. Die Wahrscheinlichkeit $1-B_k$ sinkt bei wachsendem Angebot A und bei steigendem Prioritätsklassenindex k. Im folgenden Beispiel dürfen bei n=2 Funktionseinheiten nur bis zu s=2 Anforderungen gleichzeitig warten. Stets werden die Anforderungen höchster Priorität bevorzugt zum Warten zugelassen. Die Prioritäteneinteilung entspricht wieder dem Beispiel aus Abschnitt 3.4. /17,21/

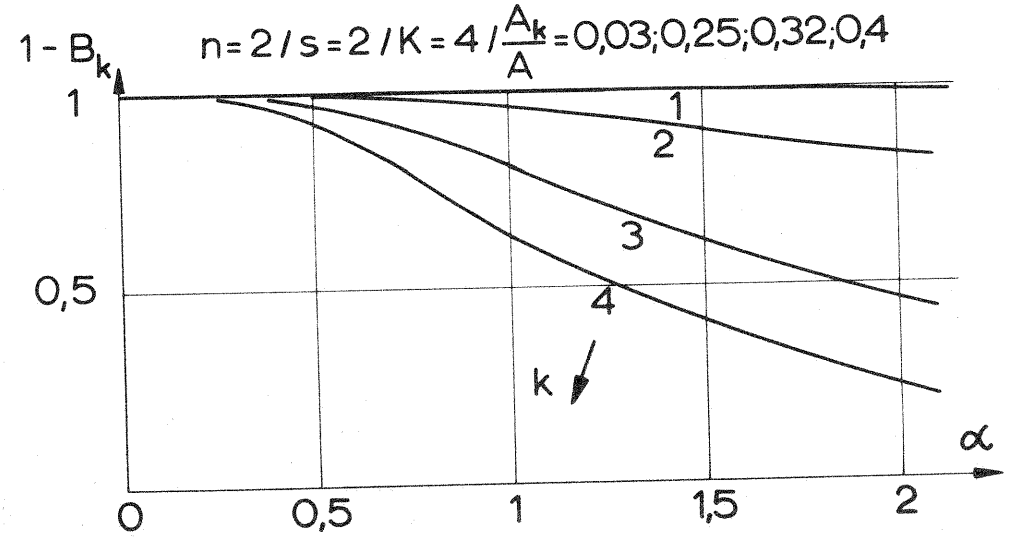

Bild 8. Die Wahrscheinlichkeit $1-B_k$, daß eine Anforderung der Prioritätsklasse k abgefertigt wird, hängt von dem auf eine Funktionseinheit bezogenen Gesamtangebot α ab. n=2 Funktionseinheiten, höchstens s=2 Anforderungen warten gleichzeitig.

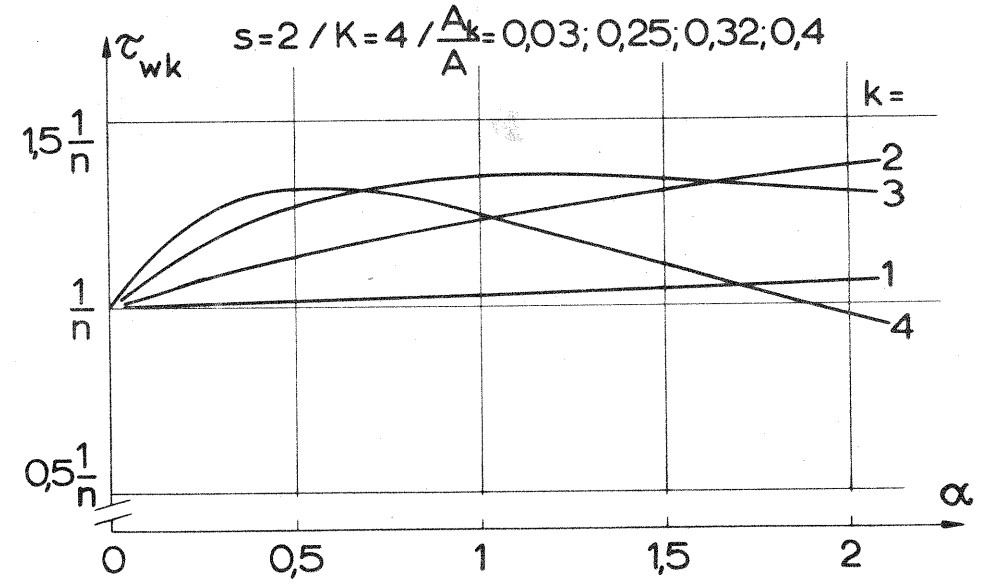

Bild 9. Mittlere Wartezeit τ_{wk} bis zum Beginn des Abfertigens.

7. Zusammenfassung

Für zwei Typen von Wartesystemen mit Prioritäten (1: Poisson-Rufprozesse, negativ exponentielle Verteilung der Belegungsdauern, $n \geq 1$ Funktionseinheiten (server) und 2: Poisson-Rufprozesse, beliebige Verteilung der Belegungsdauern, n=1) werden Wartewahrscheinlichkeit und mittlere Wartezeit der Anforderungen der Prioritätsklasse k angegeben. Für jedes beliebige Einteilen der Anforderungen in Prioritätsklassen gibt es eine obere und eine untere Grenze der mittleren Wartezeiten. Beispiele zeigen den Einfluß des Einteilens in Prioritätsklassen. Für beliebige Verteilung der Belegungsdauern und n>1 Funktionseinheiten werden Simulationsergebnisse mitgeteilt. Für die Wahrscheinlichkeit für Überschreiten bestimmter Wartedauern im ersten Wartesystemtyp wird eine einfache Näherung angegeben. Abschließend werden einige andere Arten, wie Prioritäten wirksam werden können, aufgezählt.

In diesem Beitrag sollte gezeigt werden, wie das Einteilen in Prioritätsklassen helfen kann, die Wartezeiten bei Rechensystemen zu vermindern. Das System kann dann an die von außerhalb vorgegebene unterschiedliche Wichtigkeit der Anforderungen angepaßt werden.

(für dringende Anforderungen

SCHRIFTTUM

/1/ BROCKMEYER, E., HALSTRØM, H. L. und JENSEN, A., The life and works of A. K. Erlang. Acta Polytech. Scandinavica, Copenhagen, 1960.

/2/ COBHAM, A., Priority assignment in waiting line problems. Journ. Operations Research Soc. Am. 2 (1954), 70 - 76 und 3 (1955), 547.

/3/ COFFMAN, E. G. und WOOD, R. C., Interarrival statistics for time sharing systems. Communications ACM 9 (1966), 500 - 503.

/4/ COFFMAN, E. G. und KLEINROCK, L., Some feedback queueing models for time-shared systems. Preprints of Technical Papers, Fifth International Teletraffic Congress, New York, June 14 -20, 1967, Seiten 288 - 304.

/5/ COX, D. R. und SMITH, W. L., Queues. Methuen, London, J. Wiley, New York, 1961.

/6/ DAVIS, R. H., Waiting-time distribution of a multiserver, priority queueing system. Journ. Operations Research Soc. Am. 14 (1966), 133 - 136.

/7/ DRESSIN, S. A. und REICH, E., Priority assignment on a waiting line. Quarterly of applied Math. 15 (1957), 208 - 211.

/8/ ERLANG, A. K., Lösung einiger Probleme der Wahrscheinlichkeitsrechnung von Bedeutung für die selbsttätigen Fernsprechämter. Elektrotechn. Zeitschr. (1918), 504 - 508.

/9/ HOLLEY, J., Waiting line subject to priorities. Journ. Operations Research Soc. Am. 2 (1954), 341 - 343.

/10/ KENDALL, D. G., Some problems in the theory of queues. Journ. Roy. Stat. Soc. B 13 (1951), 151 - 184.

/11/ KESTEN, H. und RUNNENBURG, J. T., Priority in waiting line problems. Proc. Kon. Ned. Akad. van Weten. Ser. A 60 (1957), 312 - 336.

/12/ LEWANDOWSKI, R., Zur Theorie der Warteschlangen. Elektron. Datenverarb. 8 (1966), 149 - 160, 208 - 218, 251 - 263.

/13/ RIORDAN, J., Stochastic service systems. J. Wiley, New York, London, 1962.

/14/ STÖRMER, H., Über ein Warteproblem aus der Vermittlungstechnik. Zeitschr. angew. Math. Mech. 40 (1960), 236 - 246 und Siemens Entwicklungsber. 23 (1960), 189 - 198.

/15/ SYSKI, R., Introduction to congestion theory in telephone systems. Oliver and Boyd, Edinburgh, London, 1960.

/16/ WAGNER, H. und DIETRICH, G., Bestimmung der Verkehrsleistung von Wartesystemen durch künstlichen Fernsprechverkehr. Nachrichtentechn. Zeitschr. 17 (1964), 273 - 279.

/17/ WAGNER, W., Über ein Wartesystem mit Nachrichten verschiedener Dringlichkeitsstufen. Bericht an die Deutsche Forschungsgemeinschaft. Institut für Nachrichtenvermittlung und Datenverarbeitung der Technischen Hochschule Stuttgart, 1. Mai 1966.

/18/ ZIMMERMANN, G. O. und STÖRMER, H., Wartezeiten in Nachrichtenvermittlungen mit Speichern.
R. Oldenbourg, München, 1961.

/19/ BJÖRKLUND, M. und ELLDIN, A., A practical method of calculation for certain types of complex common control systems. Ericsson Technics 20 (1964), 1 - 75.
4. Internat. Teletraffic Congr. 1964, London, Doc. 36.
Post Office Telecommun. Journ. 1964 Special Issue, 25.

/20/ FISZ, M., Wahrscheinlichkeitsrechnung und mathematische Statistik, Deutscher Verl. der Wissenschaften Berlin, 1958.

/21/ WAGNER, W., On combined delay and loss systems with nonpre-emptive priority service. 5. Internat. Teletraffic Congr. 1967, New York, Preprints of techn. papers, 73 - 84.