

RECHNER - PERIPHERIE

Teil 3: Warteschlangen-Berechnung

Inhaltsverzeichnis

1. Einführung	1
2. Definitionen	3
3. Verteilungsfunktionen	4
4. Warteschlangen ohne Prioritäten	6
5. Warteschlangen mit Prioritäten	9
6. Aufgaben	11

1. Einführung

Die Leistungsfähigkeit der Rechner-Peripherie bestimmt in vielen Fällen die Reaktionszeit des Rechners. Die Berechnung der Leistungsfähigkeit (Zahl der Zugriffe/s, mittlere Datenrate und Transferzeit pro Datei bei der Platte oder beim „Backup“-System ...) ist daher in der Projektphase schon von großer Bedeutung.

Bei der Berechnung kann man von unterschiedlichen Annahmen ausgehen:

- **Annahme 1: idealisiert = „nur“ Bearbeitungszeit**
Der betrachtete Vorgang (Datentransfer zu einem Gerät) läuft isoliert in Rechner ab. Es erfolgt keine Behinderung durch andere Vorgänge. Die Berechnung ist relativ einfach anhand der Gerätedaten.
- **Annahme 2: Realität = Wartezeit + Bearbeitungszeit**
Im Rechner und der Rechner-Peripherie laufen in Realität viele Vorgänge quasi-parallel ab (Multiuser- oder Multitasking-Betriebssystem). Die Vorgänge behindern sich gegenseitig; z.B. muß ein Datentransfer warten weil die Platte schon belegt ist. Zu der eigentlichen Dauer des Vorgangs müssen Wartezeiten addiert werden, die - wie sich später zeigen wird - oft länger sein können als die eigentliche Bearbeitungszeit. Wenn mehrere voneinander unabhängige Vorgänge ablaufen können die Wartezeiten nur mit statistischen Methoden berechnet werden. Die **Warteschlangentheorie** liefert dem Ingenieur die Formeln für die Berechnung dieser Wartezeiten.

Beispiele:

- **Programm-Stau vor der CPU**
Ein Prozessor arbeitet im Zeitscheibenverfahren (Multitasking-System). Die Zeitscheiben (.. 10ms ..) werden den Programmen der verschiedenen Anwender zugeteilt. Die Programme stehen in einer Warteschlange vor der CPU und konkurrieren um die Prozessorzeit.
- **Auftrags-Stau vor einem Gerät**
Die Aufträge an eine Platte werden in einer Warteschlange im Platten-Controller gesammelt und warten auf die Abarbeitung. Der Warteschlangenbetrieb eines Geräte-Controllers wird von der SCSI-Schnittstelle mit Befehlen unterstützt.

- **Transfer-Stau vor dem Server im Netz**

Programme und Daten werden meist zentral auf der Platte des Servers in einem Netz verwaltet. Bei gleichzeitigem Zugriff mehrerer Teilnehmer (PCs) auf den Server kann nur eine Anforderung bedient werden; die anderen werden in die Warteschlange des Servers eingeordnet.

- **Zyklen-Stau vor einem System-Bus**

Geräte-Controller und CPUs greifen über einen System-Bus auf einen gemeinsamen Speicher zu. Bei gleichzeitigem Zugriff darf jeweils nur ein Busteilnehmer den Bus belegen; die anderen müssen warten. Die Wartezeit am Bus verlängert die Zykluszeit. Wird der Buszugriff über eine Bussteuerung (Bus-Arbiter) verwaltet, dann bilden die anstehenden Anforderungssignale (Bus-Requests) am Bus-Arbiter praktisch die Warteschlange.

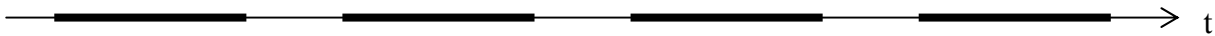
Die Formeln der **Warteschlangentheorie sind überall dort einsetzbar** wo es wegen

- unregelmäßigen Ankünften von Aufträgen und/oder wegen
 - unregelmäßigen Abfertigungszeiten von Aufträgen
- zu Stauungen kommt.

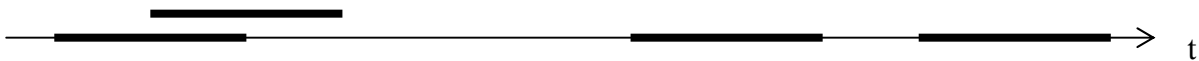
Die Warteschlangentheorie liefert die Wartezeit, die ein Auftrag „im Stau steht“ (z.B. im Stau vor dem Plattenspeicher).

Wartezeit = Zeit **von** Auftragsankunft **bis** Beginn Bearbeitung.

a) alles regelmäßig (Ankunft, Abfertigungszeiten):



b) Ankunft unregelmäßig: → Wartezeit



c) Abfertigungszeiten unregelmäßig: → Wartezeit

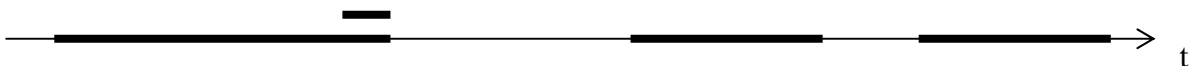


Bild: Gerät mit ca. 66% Auslastung

- a) alles regelmäßig (Ankunft, Abfertigung) → kein Stau
- b) Ankunft unregelmäßig → Stau
- c) Abfertigung unregelmäßig → Stau

2. Definitionen

Eine Warteschlange ist in einer Tabelle im Speicher, in der die wartenden Aufträge gesammelt sind, hinterlegt. Ein Auftrag für eine SCSI-Platte belegt z.B. ein 12Byte-Feld. Die Auftragsfelder sind über Zeiger verkettet; ihre Reihenfolge wird nach unterschiedlichen Prinzipien (FIFO, Priorität) geregelt. Ist das Betriebsmittel (Gerät) frei wenn ein Auftrag eintrifft, dann wird dieser sofort bearbeitet (kommt nicht in die Warteschlange). Der Anteil der wartenden Aufträge hängt von der Auslastung des Betriebsmittels ab. Nach Beendigung des Auftrags wird dieser aus der Warteschlange gelöscht.

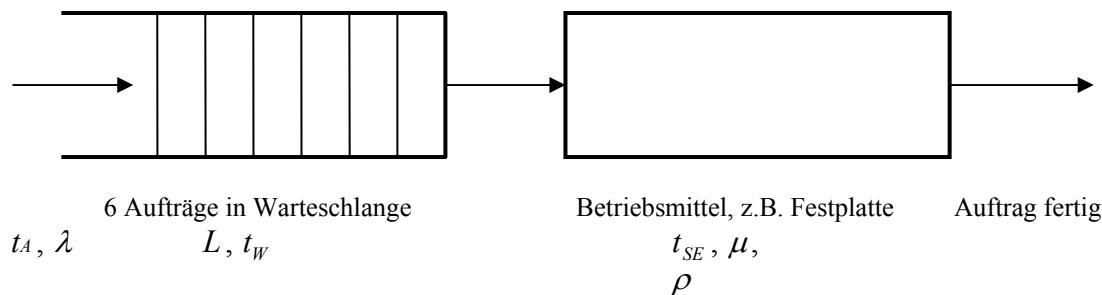


Bild: Warteschlangen-Modell

t_A	Ankunftszeit (Zeit zwischen 2 ankommenden Aufträgen)
λ	Ankunftsrate (Aufträge/s)
t_{SE}	Bedienzeit eines Auftrags (Bearbeitungszeit)
μ	Bedienrate (Aufträge/s), Abfertigungsrate
ρ	Auslastungsgrad
L	Länge der Warteschlange (Aufträge)
t_W	Wartezeit eines Auftrags in der Warteschlange
t_{EA}	Durchlaufzeit eines Auftrags

Mittelwerte: $\overline{t_A}, \overline{\lambda}, \overline{t_{SE}}, \overline{\mu}, \overline{L}, \overline{t_W}$ ($\overline{t_A}, \overline{\lambda}, \overline{t_{SE}}, \overline{\mu}, \overline{L}, \overline{t_W}$)

$$\lambda = \frac{1}{t_A} \quad \mu = \frac{1}{t_{SE}}$$

$$t_{EA} = t_W + t_{SE}$$

$$\rho = \frac{\lambda}{\mu} = \lambda \cdot t_{SE}; \quad \rho = 1 \cong 100\%, \quad \rho = 0 \cong 0\%$$

$\rho < 1$ (<100%) ist Annahme bei allen Berechnungen

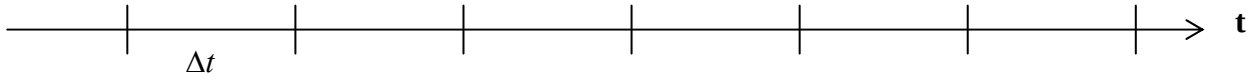
$$L = t_W \cdot \lambda = \frac{t_W}{t_A}$$

$P(t_W > 0) = \rho$ Wahrscheinlichkeit, daß ein Auftrag warten muß

$P(t_W = 0) = 1 - P(t_W > 0) = 1 - \rho$ Wahrscheinlichkeit, daß ein Auftrag nicht warten muß

3. Verteilungsfunktionen

Unter der Annahme, daß die Ankunft der Aufträge unregelmäßig ist (=statistisch unabhängig), wird ihre Verteilung durch eine Poisson-Verteilung beschrieben.



$$\bar{\lambda} * \Delta t$$

Zahl der Aufträge, die **im Mittel in** Δt eintreffen

Poisson-Verteilung:

$$P_k = \frac{(\bar{\lambda} \cdot \Delta t)^k}{k!} e^{-\bar{\lambda} \cdot \Delta t}$$

Wahrscheinlichkeit für k Aufträge in Δt

Poisson-Verteilung: beschreibt die Wahrscheinlichkeit, daß **k Aufträge in Δt** eintreffen

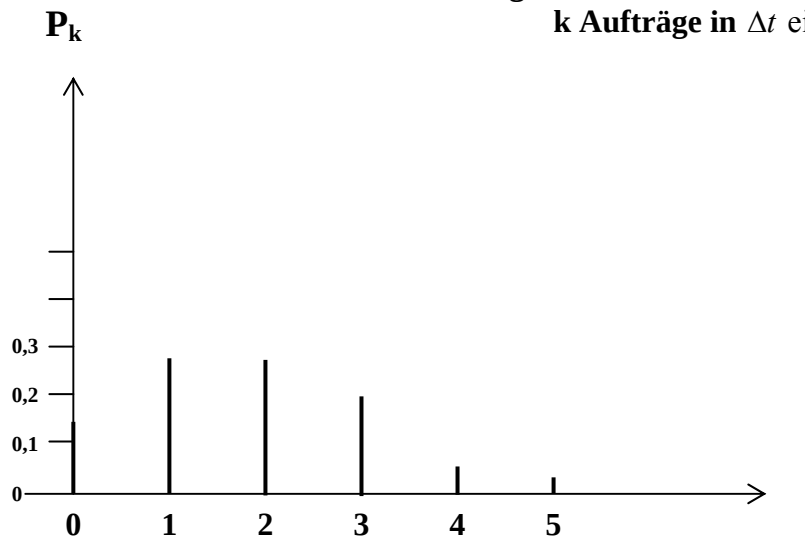
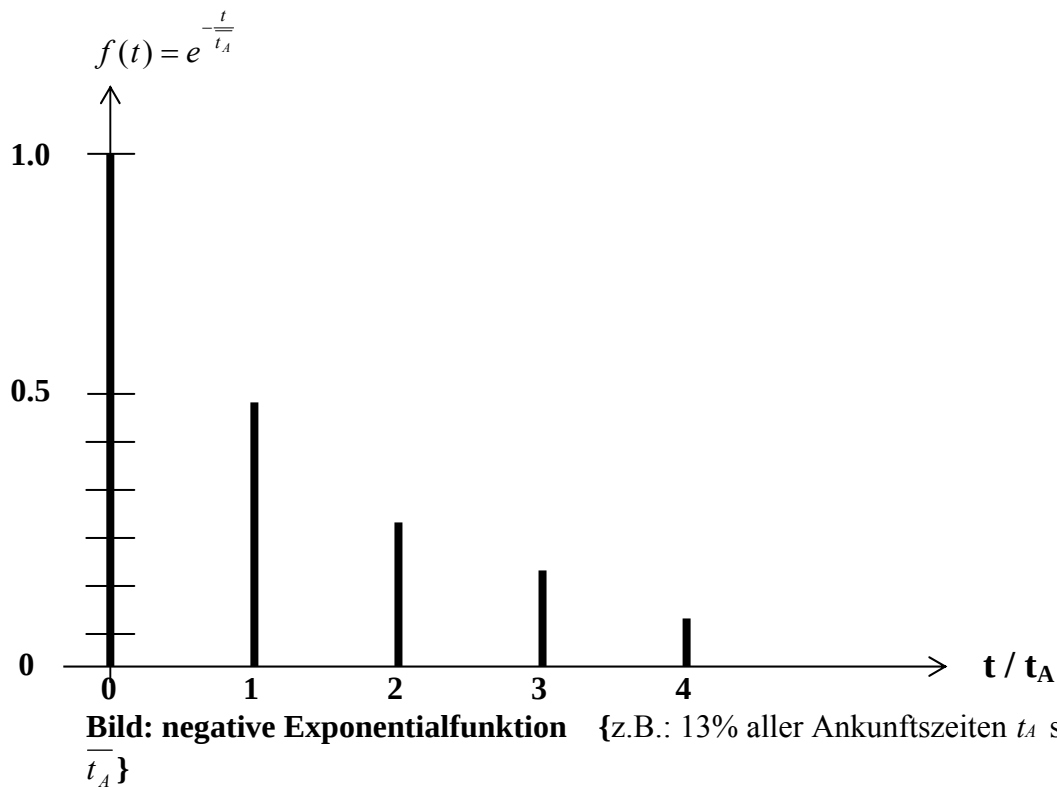


Bild: Beispiel für $\Delta t = 1\text{s}$, $\bar{\lambda} = 2$ Aufträge/s

Wenn die Aufträge nach einer Poisson-Verteilung eintreffen, dann sind die Ankunftszeiten t_A nach einer negativen Exponentialfunktion verteilt.

$$f(t) = e^{-\frac{t}{t_A}} \quad \text{mit } f(t) = \text{Wahrscheinlichkeit, daß } t_A > t$$

$$f(t) - f(t + \Delta t) \quad \text{Wahrscheinlichkeit, daß } t_A \text{ im Intervall } t < t_A \leq (t + \Delta t) \text{ liegt.}$$



$$F(t) = 1 - f(t) \quad \text{Wahrscheinlichkeit, daß } t_A \leq t.$$

Die negative Exponentialfunktion kann in gleicher Weise auch auf die Verteilung der Bedienzeiten t_{SE} angewendet werden (falls die Bedienzeiten statistisch verteilt sind).

4. Warteschlangen ohne Prioritäten

a) Bedienzeiten t_{SE} mit negat. Exponent. Verteilung

Warteschlange:

Ankunftszeiten t_A : negat. Exponent. Vert.

Bedienzeiten t_{SE} : negat. Exponent. Vert.

Disziplin: FIFO

(wer zuerst kommt wird zuerst bearbeitet = alle Aufträge haben gleiche Priorität)

Mittlere Wartezeit $\overline{t_w}$ eines Auftrags, der sich in der Warteschlange befindet:

$$\overline{t_w} = \frac{\sum \text{alle_Wartezeiten}}{\sum \text{alle_wartenden_Aufträge}} = \frac{1}{1-\rho} \cdot \overline{t_{SE}}$$

Mittlere Wartezeit $\overline{t_w^*}$, bezogen auf alle Aufträge (wartende und nicht wartende):

$$\overline{t_w^*} = \frac{\sum \text{alle_Wartezeiten}}{\sum \text{alle_Aufträge}} = \frac{\rho}{1-\rho} \cdot \overline{t_{SE}}$$
$$\overline{t_w^*} = \rho \cdot \overline{t_w}$$

Wahrscheinlichkeit $P(n)$, daß genau n Aufträge in der Warteschlange stehen:

$$P(n) = (1-\rho) \cdot \rho^n \quad n=0, 1, 2, 3, \dots$$

Wahrscheinlichkeit $P(n)$, daß mindestens n Aufträge in der Warteschlange stehen:

$$P(\geq n) = \rho^n$$

$$\{ P(0) = (1-\rho) \cdot \rho^0 = 1-\rho \quad = \text{Wahrscheinlichkeit, daß Gerät nicht belegt ist} \}$$

$$\{ P(1) = (1-\rho) \cdot \rho^1 = (1-\rho) \cdot \rho = \rho - \rho^2 \}$$

ρ = WS, daß Gerät belegt ist

ρ^2 = WS, daß mindestens 2 Aufträge in der Warteschlange steht

$$\{ P(\geq 0) = \rho^0 = 1 \quad \text{Dies trifft immer zu! Bedeutet, daß die Warteschlange entweder leer ist oder} \}$$

≥ 1 Auftrag hat

b) Konstante Bedienzeiten t_{SE}

Warteschlange:

Ankunftszeiten t_A : negat. Exponent. Vert.

Bedienzeiten t_{SE} : konstant

Disziplin: FIFO

(wer zuerst kommt wird zuerst bearbeitet = alle Aufträge haben gleiche Priorität)

Mittlere Wartezeit **in** der Warteschlange:

$$\overline{t_W} = \frac{1}{2} \cdot \frac{1}{1-\rho} \cdot \overline{t_{SE}}$$

Mittlere Wartezeit, bezogen auf **alle** Aufträge:

$$\overline{t_W^*} = \rho \cdot \overline{t_W} = \frac{1}{2} \cdot \frac{\rho}{1-\rho} \cdot \overline{t_{SE}}$$

Vergleich Fall a) mit Fall b):

Bei gleicher Auslastung liefert die Formel für konstante Bedienzeiten nur halb so große Wartezeiten. Dies ist qualitativ gut zu erklären, da ein System „konstante Bedienzeiten/negativ exp. verteilte Ankunftszeiten“ sicher weniger unregelmäßig ist als ein System „negativ exp. verteilte Bedienzeiten/negativ exp. verteilte Ankunftszeiten“.

c) Rechenbeispiel

Grafische Darstellung der Wartezeiten $\overline{t_W}$ und $\overline{t_W^*}$ als Funktion der Auslastung ρ in normierter Darstellung. Es sind konstante Bedienzeiten angenommen.

$$\frac{\overline{t_W}}{t_{SE}} = \frac{1}{2(1-\rho)}$$

$$\frac{\overline{t_W^*}}{t_{SE}} = \rho \cdot \frac{\overline{t_W}}{t_{SE}}$$

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
$\frac{\overline{t_W}}{t_{SE}}$	0.5	0.55	0.62	0.71	0.84	1	1.25	1.65	2.5	5	10
$\frac{\overline{t_W^*}}{t_{SE}}$	0	0.5	0.12	0.22	0.34	0.5	0.75	1.15	2	4.5	9.5

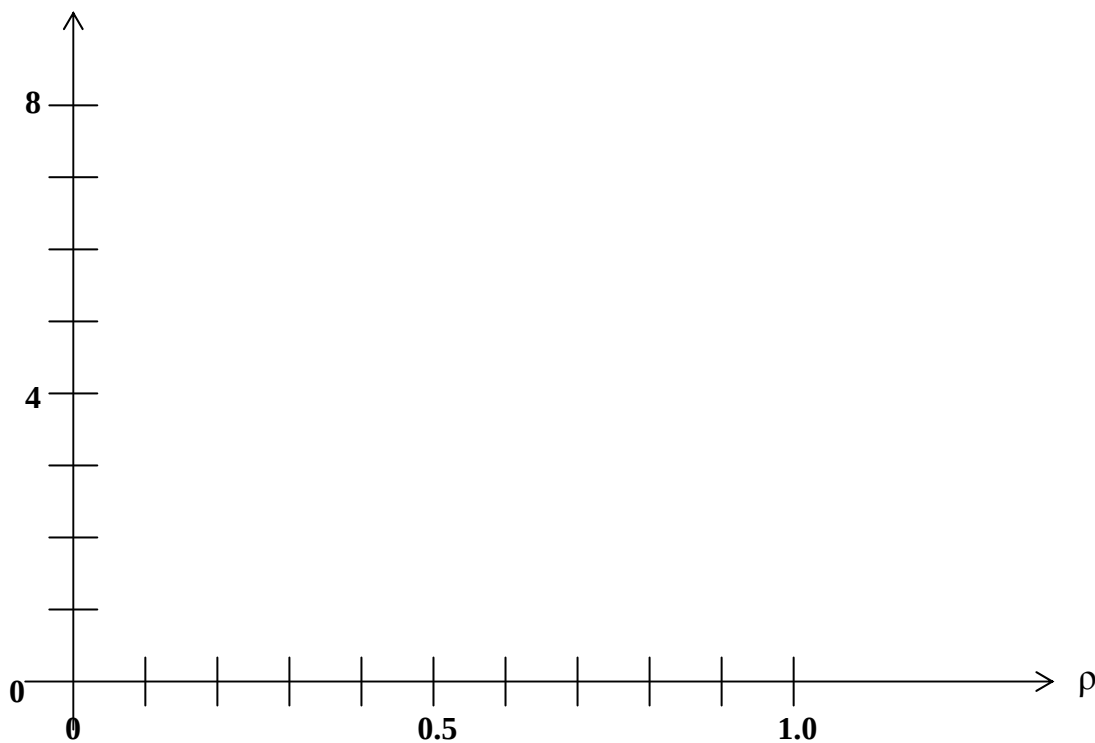


Bild:

Diskussion der Kurven:

{ $\rho=0$: ?

Höchstens 1 Auftrag in der Warteschlange und der wartet im Mittel eine halbe Bearbeitungszeit.

Die mittlere Wartezeit aller Aufträge geht gegen Null bei minimaler Auslastung.

$\rho \approx 0.6$ (=60% Auslastung):

Die mittlere Wartezeit entspricht etwa der Bearbeitungszeit.

$\rho > 0.8$ (=80% Auslastung):

Die Wartezeiten steigen steil an.}

5. Warteschlangen mit Prioritäten

Modell:

- Eine Warteschlange mit allen Aufträgen (intern geordnet nach den Prioritäten) oder formelmäßig gleichwertig:
- Für jede Priorität eine eigene Warteschlange. Abarbeitung der Warteschlangen in Reihenfolge der Prioritäten

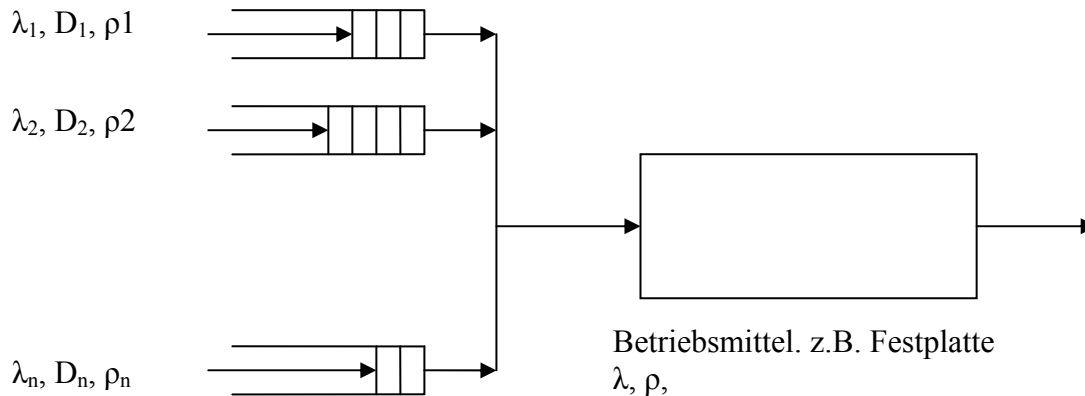


Bild: Warteschlangenmodell

λ_i Ankunftsrate der Gruppe i
 D_i Priorität (Dringlichkeit) der Gruppe i
 $\lambda = \lambda_1 + \lambda_2 + \dots = \sum \lambda_i$
 $\rho = \rho_1 + \rho_2 + \dots = \sum \rho_i$

Formelmäßig gleichwertig wäre eine einzige Warteschlange mit einer Reihenfolge der Aufträge, sortiert nach Priorität (die höchste Priorität D_1 vorne)

a) keine Unterbrechung

Warteschlange:

Ankunftszeiten t_A : negat. Exponent. Vert.

Bedienzeiten t_{SE} :
• negat. Exponent. Vert. $\rightarrow c^2=1$
• konstant $\rightarrow c^2=0$

Disziplin: Aufträge werden in der Reihenfolge ihrer **Priorität** abgearbeitet.
Keine Unterbrechnung. Höchste Priorität: D=1

Priorität D=1 (höchste Priorität):

$$\bar{t}_W(D=1) = \frac{\bar{t}_{SE}}{2(1-\rho_1)}(1+c^2)$$

$$\bar{t}_W^*(D=1) = P(t_W > 0) \cdot \bar{t}_W(D=1) = \rho \cdot \bar{t}_W(D=1)$$

Für beliebige Prioritäten:

$$\bar{t}_W(D) = \frac{\bar{t}_{SE}}{2(1-\sum_{i=1}^{D-1} \rho_i)(1-\sum_{i=1}^D \rho_i)}(1+c^2)$$

$$\bar{t}_W^*(D) = P(t_W > 0) \cdot \bar{t}_W(D) = \rho \cdot \bar{t}_W(D)$$

$$\begin{aligned} \bar{t}_{SE} &= \text{mittlere Bedienzeit, gemittelt über alle Prioritätsgruppen.} \\ &= \frac{1}{\lambda} \cdot [\lambda_1 \cdot \bar{t}_{SE}(1) + \lambda_2 \cdot \bar{t}_{SE}(2) + \lambda_3 \cdot \bar{t}_{SE}(3) \dots] \end{aligned}$$

Höchste Priorität (D=1=D₁): kürzeste Wartezeit

Niedrigste Priorität: längste Wartezeit

Beispiel:

Gruppe 1 bis 4 gegeben mit λ_i und D_i . Die Abfertigungsrate sei μ .

Gesucht $\bar{t}_W(3) = ?$

b) mit Unterbrechung

Warteschlange:

Ankunftszeiten t_A : negat. Exponent. Vert.

Bedienzeiten t_{SE} :
 • negat. Exponent. Vert. $\rightarrow c^2=1$
 • konstant $\rightarrow c^2=0$

Disziplin: Aufträge werden in der Reihenfolge ihrer **Priorität** abgearbeitet.
Mit Unterbrechung. Die Bearbeitung eines Auftrags wird durch einen ankommenden Auftrag höherer Priorität unterbrochen.

Priorität $D=1$ (höchste Priorität):

$$\bar{t}_w(D=1) = \frac{t_{SE}(1)}{2(1-\rho_1)}(1+c^2)$$

$$\bar{t}_w^*(D=1) = P(t_w > 0) \cdot \bar{t}_w(D=1) = \rho_1 \cdot \bar{t}_w(D=1)$$

6. Aufgaben

Aufgabe 1: Warteschlangen ohne Priorität

Ein Prozessor schickt Aufträge an die Festplatte. Ankunftszeiten nach negativer Exponentialverteilung.
 Der Platten-Controller verwaltet die Warteschlange für die Aufträge (keine Priorität).

Prozessor $\Rightarrow$ **Controller** $\Rightarrow$ **Festplatte**
 80% Auslastung
 200ms mittl. Bedienzeit

1. Wieviel Aufträge/s erhält der Platten-Controller?
2. Wie oft ist bei 1000 eintreffenden Aufträgen der Abstand zwischen zwei Aufträgen größer als 1 Sekunde?
3. Wie groß ist die mittlere Durchlaufzeit eines Auftrags, der in der Warteschlange steht?
4. Wie viele Aufträge stehen im Mittel in der Warteschlange?
5. Der Controller verteilt die Aufträge gleichmäßig auf zwei Plattenspeicher. Um wieviel % wird dadurch die Durchlaufzeit gegenüber Frage 3. reduziert?

Aufgabe 2: Warteschlangen mit nicht unterbrechender Priorität

Ein Prozessor bearbeitet die Programme mit einer konstanten Bedienzeit von 10 ms. Die zur Bearbeitung anstehenden Programme sind in 4 Gruppen eingeteilt.

Gruppe	A:	$\lambda_A = 10$ (1/s)	Priorität:	2	nicht unterbrechend!
	B:	$\lambda_B = 20$ (1/s)		3	
	C:	$\lambda_C = 10$ (1/s)		1	
	D:	$\lambda_D = 10$ (1/s)		3	

1. Mittlere Wartezeit der Programme aus Gruppe B und C ?
2. Wieviel Programme (1/s) aus Gruppe B und C werden ohne Wartezeit sofort bearbeitet?
3. Wie lange dauert die Abarbeitung (Durchlaufzeit) eines Programms aus Gruppe C?

4. Die Ankunftsrate $\bar{\lambda}_C$ der Programme aus Gruppe C soll variabel sein. Zeichnen Sie die mittlere Wartezeit der Programme aus Gruppe C als Funktion des Auslastungsgrads ρ_C
(Vorgehensweise:
Mittlere Wartezeit berechnen für $\rho_C = 0/0.2/$
Maßstab :
x-Achse: ρ_C , 1 cm = 0.1
y-Achse: mittlere Wartezeit, 1 cm = 1 ms)
5. Die mittlere Durchlaufzeit eines Programms aus Gruppe C soll 18ms nicht überschreiten. Entnehmen Sie aus der Zeichnung die maximal zulässige Auslastung $\rho_{C \max}$.
6. Wieso geht die mittlere Wartezeit in der Zeichnung für $\rho_C = 0$ nicht auch gegen 0 ? (Begründung stichpunktartig)

Aufgabe 3: Warteschlangen mit unterbrechender Priorität

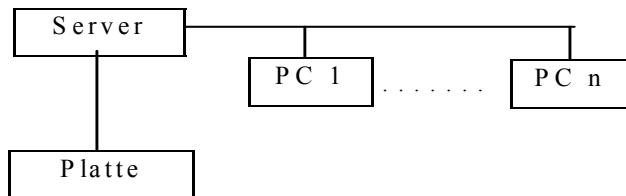
Ein Prozessor arbeitet Programme mit einer mittleren Bedienzeit $\bar{t}_{se} = 10 \text{ ms}$ ab. Die Bedienzeiten sind nach einer negativen Exponentialfunktion verteilt.
Die Programme besitzen verschiedene Prioritäten, höherpriorige unterbrechen niederpriorige Programme.
Warteschlangenbetrieb.

$$\begin{array}{llll} \bar{\lambda}_A & = & 10 \text{ (1/s)} & D = 1 \\ \bar{\lambda}_B & = & 40 \text{ (1/s)} & D = 2 \\ \bar{\lambda}_C & = & 30 \text{ (1/s)} & D = 3 \end{array} \quad (\bar{\lambda} = \text{mittlere Ankunftsrate, } D = \text{Priorität})$$

1. Wieviel % der Programme haben Bedienzeiten, die größer sind als
 - a) die mittlere Bedienzeit $\bar{t}_{se}$.
 - b) 50 ms.
2. Wie lange dauert im Mittel die Abarbeitung (Durchlaufzeit) eines Programms mit der höchsten Priorität?
3. Die Programme werden auf 2 Prozessoren gleichmäßig verteilt. Die Bedienzeiten sind 20 % höher, da beide Prozessoren auf einen gemeinsamen Speicher zugreifen.
 - a) Wie hoch ist die Auslastung (in %) der Prozessoren?
 - b) Wie lange dauert im Mittel die Abarbeitung (Durchlaufzeit) eines Programms mit der höchsten Priorität?

Aufgabe 4: Warteschlangen ohne Priorität

Im Labor sind mehrere PCs über ein lokales Netz an einen Zentralrechner(Server) angeschlossen. Programme und Daten der PCs sind auf der Platte des Servers abgespeichert und werden bei Bedarf von den PCs angefordert. Der Server verwaltet die Aufträge für die Platte (Warteschlange). Ankunftsrate und Bedienrate angenähert durch negative Exponentialverteilung.



$$\bar{t}_p = 10 \text{ ms}$$
$$5400 \text{ U/min}$$

1. Ankunftsrate pro PC: $A=120$ (Aufträge/min). Pro Auftrag sind im Mittel 2 Spuren (auf dem **gleichen** Zylinder) zu lesen (oder zu schreiben). Berechnen Sie die Auslastung (%) des Servers als Funktion von n (n = Zahl der PC).
2. a) Wieviele (n_1) PC sind maximal anschließbar
b) Wieviele (n_2) PC sind maximal anschließbar, wenn die mittlere Länge der Warteschlange 5 Aufträge nicht überschreiten soll?
3. Die Hälfte der Aufträge sind Schreib-Aufträge. Nach jedem Schreiben einer Spur wird jetzt automatisch ein Kontroll-Lesen der Spur durchgeführt. Berechnen Sie die mittlere Bedienzeit (t_{se}) des Servers.
4. a) Die Auslastung des Servers durch einen PC beträgt jetzt 10%. Ankunftsrate pro PC: $A=120$ (Aufträge/min). Berechnen Sie die Bedienzeit t_{se} .
b) Summieren Sie die Wartezeiten (t_{sum}), die ein PC während einer Stunde Betriebszeit am Server wartet, wenn am Server 6 PC angeschlossen sind.

-----**Ab hier: Semester-Ferien**-----