Warteschlangentheorie

Klaus Berger
Pantelis Christodoulides
Karl Grill[1]

Version 1.0
October 11, 1999

Inhaltsverzeichnis

Kapitel 1

Einleitung

Wir betrachten folgendes grundlegendes Modell: Kunden kommen zu zufälligen Zeiten $T_1 < T_2 < \ldots < T_n < \ldots$ im System an, wobei T_n die Ankunftszeit des n-ten Kunden bezeichnet.

Ein oder mehrere Bediener arbeiten die Schlange ab und für jeden Kunden wird eine bestimmte „Bedienzeit“ benötigt. Es sei x_n die Bedienzeit des n-ten Kunden. Die Reihenfolge der Bedienung der Kunden wird durch die sogenannte „Disziplin“ der Warteschlange bestimmt.Wir nehmen meistens FCFS (First Come First Serve) an. Andere Möglichkeiten wären LCFS (Last Come First Serve) oder „Prioritäten“.

Folgende Annahmen werden getroffen:

1. Die x_n sollen unabhängig und identisch verteilt sein.
2. t_n ist die n-te Zwischenankunftszeit also $t_n = T_n - T_{n-1}$, $T_0 = 0$ (Die Zeit zwischen der Ankunft des n-ten und des $(n-1)$-ten Kunden). Die t_n sind auch unabhängig und identisch verteilt.

Wir verwenden folgende Kurznotation für Warteschlangen: $A/B/s$.

$A \ldots$ Verteilung der Zwischenankuftszeiten t_n, wobei a die Dichte von t_n ist.
$B \ldots$ Verteilung der Bedienzeiten x_n wobei b die Dichte von x_n ist.
$s \ldots$ Anzahl der Bediener (Server).

Kurznotationen für Verteilungen sind:

M ... Exponentialverteilung („memoryless“).
Dichtefunktion:

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0.$$

E_n ... Erlangverteilung: Die Summe von n unabhängigen Exponentialverteilungen.
Dichtefunktion:

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

H ... Hyperexponentielle Verteilung: Die Mischung von unabhängigen Exponentialverteilungen. Wir haben $p_1 \dots p_n$, $p_i \geq 0$, und $\sum_{i=1}^{n} p_i = 1$, $\lambda_1 \dots \lambda_n \geq 0$.
Dichtefunktion:

$$f(x) = \sum_{i=1}^{n} p_i \lambda_i e^{-\lambda_i x}.$$

D ... Deterministisch: Ein fixer Wert wird angenommen.
G ... „General“: Allgemeine Verteilung (alles, was nicht vorher erwähnt wurde).

Die Sonderstellung der Exponentialverteilung ist begründet durch ihre Gedächtnislosigkeit. Falls nämlich etwa eine Wartezeit exponentialverteilt ist, und wir schon t Zeiteinheiten gewartet haben, so ist die Verteilung der restlichen Wartezeit gegeben durch

$$\begin{aligned} &\mathbb{P}(\text{restliche Wartezeit} \geq x \mid \text{schon } t \text{ gewartet}) = \\ &= \mathbb{P}(T \geq t + x \mid T \geq t) = \frac{\mathbb{P}(T \geq t + x)}{\mathbb{P}(T \geq t)} \ . \end{aligned}$$

Angenommen T sei exponentialverteilt $\Rightarrow \mathbb{P}(T \geq t) = e^{-\lambda t} \Rightarrow$

$$\frac{\mathbb{P}(T \geq t + x)}{\mathbb{P}(T \geq t)} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x},$$

also unabhängig davon, wie lange wir schon vorher gewartet haben.

Es gibt abgeleitete Größen, die das Verhalten der Warteschlange beschreiben wie:

1. w_n ... Wartezeit des n-ten Kunden.

2. $z_n = w_n + x_n \ldots$ Zeit, die der n-te Kunde im System verbringt.
3. N_t ... Anzahl der Kunden, die zum Zeitpunkt t im System sind (= wartende + eventuell die, die gerade bedient werden).

Es gibt einige Fragen, die uns interessieren:

1. Die Verteilungen von w_n, z_n, N_t.
2. Gibt es Grenzverteilungen für $n \to \infty$ bzw. $t \to \infty$ (d.h. pendelt sich das Verhalten der Schlange auf einen stationären Zustand ein?) und Bestimmung der Grenzverteilungen.
3. Erwartungswerte der Grenzverteilungen in 2.
4. Abschätzungen für 3.

Die Aufgaben sind hier in abnehmender Schwierigkeit geordnet. Leider sind die genauen Verteilungen 1. nicht leicht zu bestimmen, also beschränken wir uns meist auf 2. ; im ganz allgemeinen Fall wird es sogar notwendig sein, nur Abschätzungen zu betrachten.

Kapitel 2

Erste Resultate

2.1 Eine Rekursion für die Wartezeit

Wir wollen nun die Wartezeit des $(n+1)$-ten Kunden durch die des n-ten Kunden ausdrucken. Dazu ist

1. $T_n \ldots$ die Ankunftszeit des n-ten Kunden.
2. $T_n + w_n \ldots$ die Zeit, wenn der n-te Kunde bedient wird.
3. $T_n + w_n + x_n \ldots$ die Zeit wenn der n-te Kunde geht. Ab jetzt kann der $(n+1)$-te bedient werden.
4. $T_{n+1} = T_n + t_{n+1} \ldots$ Ankuftszeit des $(n+1)$-ten Kunden.

Falls $T_{n+1} < T_n + w_n + x_n$, dann ist $w_{n+1} = T_{n+1} + w_n + x_n - T_{n+1} = w_n + x_n - t_{n+1}$. Falls $T_{n+1} \geq T_n + w_n + x_n$ ist $w_{n+1} = 0$. Also

$$w_{n+1} = \max(w_n + x_n - t_{n+1}, 0) =: (w_n + x_n - t_{n+1})_+ \ .$$

Sei $u_n = x_n - t_{n+1}$. Die u_i sind unabhängig und identisch verteilt.

$$\begin{aligned} \Rightarrow w_n &= \max(w_{n-1} + u_{n-1}, 0) = 0 \\ \Rightarrow w_n &= \max(0, u_{n-1} + \max(w_{n-2} + u_{n-2}, 0)) = \\ &= \max(0, u_{n-1}, u_{n-1} + u_{n-2} + w_{n-2}) = \ldots \\ &= \max(0, u_{n-1}, u_{n-1} + u_{n-2}, \cdots, u_{n-1} + u_{n-2} + \cdots + u_1) \ . \end{aligned}$$

Also ist die Verteilung von w_n dieselbe wie die von $\tilde{w}_n$ mit

$$\tilde{w}_n = \max(0, u_1, u_1 + u_2, \cdots, u_{n-1} + u_{n-2} + \cdots + u_1) \ .$$

Offensichtlich ist $\tilde{w}_n$ eine monoton nichtfallende Folge, also existiert

$$\tilde{w} = \lim_{n\to\infty} w_n.$$

Falls $\mathbb{E}u > 0$, dann geht $u_1 + \cdots + u_{n-1} \to \infty$, also auch $\tilde{w}$. Falls $\mathbb{E}u < 0$, dann geht $u_1 + \cdots + u_{n-1} \to -\infty$, also ist für $n > n_0$ $u_1 + \cdots + u_{n-1} < 0$, was bedeutet daß nur die ersten Glieder in der Definition von $\tilde{w}_n$ wichtig sind; also ist $\tilde{w}$ endlich. Falls $\mathbb{E}u = 0$, ist können wir den einfachen Fall $D/D/1$ betrachten. In diesem Fall ist $w_n = 0$, also ist das Verhalten stationär. Leider ist das der einzige Fall; sobald A oder B nicht degenerierte Verteilungen haben, kann $\tilde{w}_n$ nicht gegen eine endliche Zufallsvariable konvergieren, weil etwa nach dem zentralen Grenzwertsatz für n groß genug

$$\mathbb{P}(\tilde{w}_n > a\sqrt{n.\mathbf{Var}(u)}) \geq 1 - \Phi(a) - \epsilon > 0 \ .$$

Somit ist für jedes n

$$\mathbb{P}(\tilde{w} > a\sqrt{n.\mathbf{Var}(u)}) \geq 1 - \Phi(a) - \epsilon \ ,$$

also

$$\mathbb{P}(\tilde{w} = \infty) \geq 1 - \Phi(a) - \epsilon \ .$$

Es bleibt uns also für stationäres Verhalten (außer im Trivialfall $D/D/1$) die Bedingung

$$\mathbb{E}(u) < 0 \Leftrightarrow \mathbb{E}(x) < \mathbb{E}(t) \Leftrightarrow \rho = \frac{\mathbb{E}(x)}{\mathbb{E}(t)} < 1 \ .$$

Wir bezeichnen den Kehrwert von $\mathbb{E}(t)$ als die „Ankunftsrate“ $\lambda = \frac{1}{\mathbf{E}(t)}$ und $\mu = \frac{1}{\mathbf{E}(x)}$ als die „Bedienrate“. Es sind dies die Anzahl von Kunden, die in einem langen Zeitraum durchschnittlich pro Zeiteinheit ankommen bzw. bedient werden, falls ununterbrochen bedient wird. Mit diesen Bezeichnungen ist $\rho = \frac{\lambda}{\mu}$ (Auslastung) und die Bedingung für stationäres Verhalten wird zu $\lambda < \mu$ bzw. $\rho < 1$.

2.2 Der Satz von Little

Wir nehmen jetzt an, daß in unserer Schlange stationäres Verhalten herrscht; wir wollen eine Beziehung zwischen der Ankuftsrate, der mittleren Anzahl der Kunden im System und der mittleren Aufenthaltsdauer finden. Dazu nehmen wir an, daß wir jeden Kunden für die Zeit, die er im System verbringt, bezahlen müssen. Die Summe, die insgesamt zu bezahlen ist, berechnet sich als $T\mathbb{E}(N)$, da zu jedem Zeitpunkt durchschnittlich $\mathbb{E}(N)$ Kunden anwesend sind. Andererseits bekommt jeder Kunde durchschnittlich $\mathbb{E}(z)$ bezahlt. In der Zeit T kommen λT Kunden an, also ist die zu bezahlende Summe auch gleich $\lambda T\mathbb{E}(z)$.

Beide Gleichungen sind nicht vollständig exakt, weil in beiden Fällen noch zufällige Schwankungen dazukommen, und weil bei der zweiten Gleichung auch nicht berücksichtigt wurde, daß einige Kunden noch nach T bleiben. Diese Fehler sind aber von kleineren Ordnung als T. Wir haben also

$$T\mathbb{E}(N) = \lambda T\mathbb{E}(z) + o(T) \ .$$

Dividieren wir durch T und $T \to \infty$ gibt

$$\mathbb{E}(N) = \lambda\mathbb{E}(z) \ ,$$

d.h. Mittlere Anzahl = Ankuftsrate $*$ Mittlere Aufenthaltsdauer. Wendet man dieses Ergebnis auf den Server allein an, ergibt sich, daß die Mittlere Anzahl der Kunden, die gerade bedient werden =

$$\lambda\mathbb{E}(x) = \frac{\lambda}{\mu} = \rho \ .$$

Da aber höchstens 1 Kunde bedient wird, ist das gleich der Wahrscheinlichkeit, daß der Server besetzt ist, oder der Auslastung des Servers.

Kapitel 3

Warteschlangensysteme

3.1 Die Schlange $M/M/1$

Im Folgenden gehen wir von der „FCFS“-Disziplin aus. Um die zukünftige Entwicklung einer Warteschlange bestimmen zu können, benötigen wir 3 Angaben zur Zeit t.

1. Die Anzahl N_t der anwesenden Kunden.
2. Die Zeit, die seit der letzten Ankunft vergangen ist.
3. Die Zeit, die seit dem Beginn des letzten Bedienvorgangs vergangen ist (falls dieser noch andauert).

Die letzten beiden Angaben sind notwendig, damit wir die Verteilung der verbleibenden Zeit bis zur nächsten Ankunft bzw. bis zum Ende des Bedienvorganges bestimmen können. Für den Fall $M/M/1$ sind diese Angaben nicht notwendig, weil diese Verteilungen wegen der Gedächtnislosigkeit der Exponentialverteilung nicht von der schon verstrichenen Zeit abhängen. Deshalb genügt uns N_t zur Beschreibung des Systems.

Wir betrachten jetzt die Anzahl der Kunden zur Zeit $t+\Delta t$, wenn die Anzahl zur Zeit t bekannt ist. Die Anzahl kann sich in folgender Weise ändern:

1. Es kann gar nichts geschehen.

2. Es kann genau ein Kunde aufkommen.

3. Es kann genau ein Kunde fertig werden.

4. Es kann mehr als ein Ereignis (Ankunft, gehen) auftreten.

Die Wahrscheinlichkeit, daß mindestens ein Kunde im Intervall $(t, t+\Delta t)$ ankommt, ist $1 - e^{-\lambda \Delta t} = \lambda \Delta t + o(\Delta t)$. Ebenso ist die Wahrscheinlichkeit, daß ein Kunde fertig wird $\mu \Delta t + o(\Delta t)$. Die Wahrscheinlichkeit für 4. ist, wie man leicht einsieht, $o(\Delta t)$. Das gibt für 1. die Wahrscheinlichkeit $1 - (\lambda + \mu)\Delta t + o(\Delta t)$. Falls die Schlange leer ist, fallen natürlich die Summanden mit μ weg (es kann ja niemand gehen). Somit gilt für

$$\begin{aligned} p_n(t) &= \mathbb{P}(N_t = n) \\ p_n(t+\Delta t) &= \mu \Delta t . p_{n+1}(t) + (1 - (\lambda+\mu)\Delta t) p_n(t) + \\ &\quad + \lambda \Delta t p_{n-1}(t) + o(\Delta t) \qquad [n \geq 1] \\ p_0(t+\Delta t) &= \mu \Delta t . p_1(t) + (1 - \lambda \Delta t) p_0(t) + o(\Delta t) \ . \end{aligned}$$

Wenn man $p_n(t)$ auf die linke Seite bringt und durch Δt dividiert und $\Delta t \to 0$ läßt, ergibt sich

$$\begin{aligned} p_n'(t) &= \mu p_{n+1}(t) - (\lambda + \mu) p_n(t) + \lambda p_{n-1}(t) \\ p_0'(t) &= \mu p_1(t) - \lambda p_0(t) \ . \end{aligned}$$

Diese Gleichungen lassen sich etwa mit Hilfe von Transformationen lösen, aber das Ergebnis ist nicht besonders schön. Wir beschränken uns daher jetzt und in der Folge auf die Bestimmung der stationären Lösung. Diese ist natürlich dadurch gekennzeichnet, daß $p_n(t)$ nicht von der Zeit t abhängt, also $p_n'(t) = 0$. Das ergibt die Gleichungen

$$\begin{aligned} \mu p_{n+1} - (\lambda + \mu) p_n + \lambda p_{n-1} &= 0 \\ \mu p_1 - \lambda p_0 &= 0 \ . \end{aligned}$$

Durch Induktion erhalten wir daraus

$$\mu p_{n+1} - \lambda p_n = 0 \ ,$$

oder

$$p_{n+1} = \frac{\lambda}{\mu} p_n = \rho p_n \ .$$

Also ist

$$p_n = \rho^n p_0$$

und wegen $\sum_{n=0}^{\infty} p_n = 1$

$$p_0 = 1 - \rho$$

und

$$p_n = \rho^n (1 - \rho) \ .$$

Die Anzahl der Kunden im System ist also geometrisch verteilt. Aus dieser Verteilung können wir jetzt die Verteilungen von w und z bestimmen. Die Zeit im System z, falls bei der Ankunft n Personen anwesend sind, ist die Summe von $(n+1)$ exponentialverteilten Zufallsvariablen (die Bedienzeiten der n anwesenden + die der neu hinzugekommenen), hat also die Dichte

$$f_z(u|N_t = n) = \frac{u^n}{n!} \mu^{n+1} e^{-\mu u} \qquad [u > 0] \ .$$

Die unbedingte Dichte ergibt sich also zu

$$\begin{aligned} f_z(u) &= \sum_n \mathbb{P}(N_t = n) . f_z(u|N_t = n) = \\ &= \sum_n (1-\rho)\rho^n . \frac{u^n}{n!} \mu^{n+1} e^{-\mu u} = \\ &= (1-\rho) e^{-\mu(1-\rho)u} \qquad [u > 0] \ . \end{aligned}$$

z ist also exponentialverteilt mit Parameter

$$\mu(1 - \rho) = \mu - \lambda \ .$$

Die Verteilung von w ist gemischt: $\mathbb{P}(w = 0) = 1 - \rho$, und die bedingte Verteilung von w unter der Bedingung $[w > 0]$ ist wieder eine Exponentialverteilung mit Parameter $\mu - \lambda$.

3.2 Das System $M/G/1$

Jetzt benötigen wir zusätzlich zu N_t die Information über die schon verbrauchte Bedienzeit. Die einfachste Methode besteht darin, das System nur in solchen Zeitpunkten zu betrachten, an denen die verbrauchte Bedienzeit

bekannt ist (und zwar = 0), nämlich die Zeitpunkte T_n, in denen der n-te Kunde das System verläßt. Es sei N_n die Anzahl der Kunden, die dann im System verbleiben. Es gilt: Falls $N_n = 0$, so muß zuerst gewartet werden, bis ein neuer Kunde ankommt; wenn dieser Kunde geht, sind noch genau die Kunden da, die während seiner Bedienzeit angekommen sind; bezeichnet man M_n als die Anzahl der Kunden, die während der Bedienzeit des n-ten Kunden ankommen, so gilt

$$N_{n+1} = M_n \ .$$

Falls $N_n \neq 0$ ist

$$N_{n+1} = N_n - 1 + M_n \ .$$

Zusammengefaßt ergibt sich:

$$N_{n+1} = (N_n - 1)_+ + M_n \ .$$

Wir suchen eine stationäre Lösung; es sei also $\mathbb{P}(N_n = k) = p_k$ unabhängig von n. Die erzeugende Funktion von N_n ist

$$P^*(z) = \sum p_k z^k \ .$$

Die erzeugende Funktion von $(N_n - 1)_+ =$

$$\begin{aligned} = p_0 + \sum_{k=1}^{\infty} p_k z^{k-1} &= p_0 + \frac{\hat{P}(z) - p_0}{z} = \\ &= \frac{\hat{P}(z) - p_0(1-z)}{z} \ . \end{aligned}$$

Mithilfe der Transformationen (Anhang A) ergibt sich die erzeugende Funktion von M_n (die Ankünfte bilden ja einen Poisson - Prozeß) als

$$\tilde{B}(\lambda(1-z)) \ ,$$

wobei B die Verteilung der Bedienzeit (mit Dichte β) ist. Wir erhalten also

$$P^*(z) = \frac{(P^*(z) - p_0(1-z))}{z} \tilde{B}(\lambda(1-z)) \ ,$$

also

$$P^*(z) = \frac{p_0(1-z)\tilde{B}(\lambda(1-z))}{\tilde{B}(\lambda(1-z)) - z} \ .$$

Hier ist noch p_0 zu bestimmen, und zwar aus der Bedingung $P^*(1) = 1$. Es ergibt sich $p_0 = 1 - \rho$ und

$$P^*(z) = \frac{(1-\rho)(1-z)\tilde{B}(\lambda(1-z))}{\tilde{B}(\lambda(1-z)) - z} ,$$

eine sogenannte Pollaczek - Khinchin Formel. Die Anzahl der Kunden, die der n-te Kunde zurückläßt, ist genau die Anzahl der Kunden, die ankommen während er im System ist (d.h. während z_n), d.h. für die L-Transformierte $\tilde{Z}(t)$ der Verteilung von z gilt:

$$\tilde{Z}(\lambda(1-z)) = P^*(z) ,$$

also

$$\tilde{Z}(t) = \frac{(1-\rho)t\tilde{B}(t)}{t + \lambda\tilde{B}(t) - \lambda} .$$

Auch das nennt man eine Pollaczek - Khinchin Formel. Für die Wartezeit gilt (wegen $z_n = w_n + x_n$)

$$\tilde{Z}(t) = \tilde{W}(t)\tilde{B}(t) ,$$

also

$$\tilde{W}(t) = \frac{(1-\rho)t}{t + \lambda\tilde{B}(t) - \lambda} .$$

Für die Erwartungswerte ergibt sich:

$$\begin{aligned} \mathbb{E}(N) &= \rho + \frac{\lambda^2 \mathbb{E}x^2}{2(1-\rho)} \\ \mathbb{E}(Z) &= \frac{1}{\mu} + \frac{\lambda \mathbb{E}x^2}{2(1-\rho)} \\ \mathbb{E}(W) &= \frac{\lambda \mathbb{E}x^2}{2(1-\rho)} . \end{aligned}$$

3.3 Das System $G/M/1$

Jetzt betrachten wir analog zum vorigen Kapitel das System zu den Zeiten T_n, wo der n-te Kunde ankommt. N_n sei die Anzahl der anwesenden Kunden, die der n-te Kunde vorfindet.

$$N_{n+1} = N_n + 1 - \text{Anzahl der Kunden, die während } t_{n+1} \text{ gehen.}$$

Für $n > 0$ gilt, wenn wir $p_k = \mathbb{P}(N_n = k)$ (stationär!) setzen:

$$p_k = \sum_{j=k-1}^{\infty} p_j q_{j+1-k} \qquad [k \geq 1] \ ,$$

wobei

$$q_s = \mathbb{P}(\ s \text{ Kunden gehen während } t_{n+1}) =$$

$$= \mathbb{P}(\text{während } t_{n+1} \text{ treten genau } s \text{ Ereignisse eines Poissons -}$$
$$\text{Prozesses mit Rate } \mu \text{ auf}) \ .$$

Die Gleichung für $k = 0$ ist überflüssig, da sie aus den Gleichungen für $k > 0$ und der Beziehung $\sum p_k = 1$ gefolgert werden kann. Man kann zeigen, daß diese Gleichung eine eindeutige Lösung besitzt. Falls nun (p_k) eine Lösung ist, ist auch

$$\tilde{p}_k = \frac{p_{k+1}}{1 - p_0}$$

eine Lösung. Es muß also

$$\tilde{p}_k = p_k \ ,$$

somit

$$p_{k+1} = p_k(1 - p_0)$$

und

$$p_k = p_0(1 - p_0)^k = \sigma^k(1 - \sigma) \qquad [\sigma := 1 - p_0] \ .$$

Setzt man das in die Gleichung für $k = 1$ ein, ergibt sich

$$\sigma = \sum_{j=0}^{\infty} \sigma^j q_j = \tilde{A}(\mu(1 - \sigma)) \ .$$

Falls $\rho < 1$, gibt es genau eine Lösung $\sigma \in (0, 1)$. Dann ist N geometrisch verteilt mit Parameter σ. Wie für die Schlange $M/M/1$ ergibt sich die Verteilung der Zeit z im System als Exponentialverteilt mit Parameter $\mu(1-\sigma)$; die Wartezeit w hat $\mathbb{P}(w = 0) = 1 - \sigma$ und die bedingte Verteilung von w unter $[w > 0]$ ist wieder dieselbe Exponentialverteilung wie die von z.

3.4 Das System $G/G/1$

Hier sind beide Verteilungen - die der Zwischenankunftszeiten und die der Bedienzeiten - allgemeine Verteilungen. Der Trick der vorigen beiden Kapitel funktioniert jetzt nicht mehr gut. Um beide Zeiten zu kontrollieren, müßten wir das System nun zu den Zeitpunkten betrachten, in denen ein Kunde das leere System betritt; diese Zeitpunkte sind aber zu selten, um vernüftig damit zu arbeiten. Statt dessen gehen wir von der Rekursion für die Wartezeiten aus:

$$w_{n+1} = (w_n + u_n)_+ \ .$$

Das bedeutet für die Verteilungsfunktion $W(.)$ von w

$$W(x) = \mathbb{P}(w_{n+1} \leq x) = \begin{cases} \mathbb{P}(w_n + u_n \leq x) & x \geq 0 \\ 0 & x < 0 \end{cases} \ .$$

Die Wahrscheinlichkeit $\mathbb{P}(w_n + u_n \leq x)$ berechnet sich als

$$\mathbb{P}(w_n + u_n \leq x) = \int_{-\infty}^{\infty} W(x-u)c(u)du \ ,$$

wobei $c(u)$ die Dichte von $u_n = x_n - t_{n+1}$ ist. Falls in der Gleichung für $W(x)$ die Fallunterscheidung nicht auftreten würde, wäre sie leicht durch Transformationen zu lösen. Wir erreichen dies durch einen Kunstgriff: Wir setzen

$$Y(x) = \begin{cases} \int_{-\infty}^{\infty} W(x-u)c(u)du & x < 0 \\ 0 & x \geq 0 \end{cases} \ .$$

Dann ist

$$W(x) + Y(x) = \int_{-\infty}^{\infty} W(x-u)c(u)du \ .$$

Wir bezeichnen jetzt die Laplace - Transformierte von W mit $\Phi(t)$, und die von Y mit $\Phi^-(t)$. Durch partielle Integration zeigt man, daß

$$\Phi(t) = \frac{1}{t}\tilde{W}(t)$$

gilt. Für die Transformationen ergeben sich die Formeln

$$\Phi(t) + \Phi^-(t) = \Phi(t)\tilde{C}(t) = \Phi(t)\tilde{A}(-t)\tilde{B}(t) \ ,$$

oder

$$\frac{\Phi^-(t)}{\Phi(t)} = \tilde{A}(-t)\tilde{B}(t) - 1 \ .$$

Wir nehmen an, daß $\tilde{A}(t)$ für $t \geq -D$ existiert. (Das ist gleichbedeutend damit, daß $\mathbb{P}(t_n \geq x)$ wie e^{-Dx} fällt). Dann existiert $\tilde{A}(-t)\tilde{B}(t) - 1$ für $0 \leq t \leq D$; Ferner existiert $\Phi(t)$ für $t > 0$ und $t\Phi(t)$ ist in $\Re(t) \geq 0$ regulär und beschränkt; $\Phi^-(t)$ existiert für $t \leq D$ und in $\Re(t) < D$ ist $t\Phi^-(t)$ regulär und beschränkt. Wir versuchen 2 Funktionen Ψ^+ und Ψ^- zu finden, die folgendes erfüllen:

1. $\frac{\Psi^+(t)}{\Psi^-(t)} = \tilde{A}(-t)\tilde{B}(t) - 1$ (Spektralzerlegung).
2. $\frac{\Psi^+(t)}{t}$ ist für $\Re(t) > 0$ regulär und beschränkt und hat dort keine Nullstellen.
3. $\frac{\Psi^-(t)}{t}$ ist für $\Re(t) < D$ regulär und beschränkt und hat dort keine Nullstellen.

Dann gilt

$$\frac{\Phi^-(t)}{\Phi(t)} = \frac{\Psi^+(t)}{\Psi^-(t)} \qquad 0 < \Re(t) < D \ ,$$

oder

$$\Phi^-(t)\Psi^-(t) = \Psi^+(t)\Phi(t) \qquad 0 < \Re(t) < D \ .$$

Die linke Seite ist für $\Re(t) < D$ regulär und beschränkt, die rechte Seite für $\Re(t) > 0$. Es ist dadurch also eine Funktion bestimmt, die in der ganzen Ebene regulär und beschränkt ist. Nach dem Satz von LIOUVILLE muß eine solche Funktion konstant sein. Es gilt also

$$\Phi(t) = \frac{K}{\Psi^+(t)} \ ,$$

und

$$\tilde{W}(t) = \frac{Kt}{\Psi^+(t)} \ .$$

Es bleibt die Konstante K zu bestimmen. Sie folgt wieder aus

$$\tilde{W}(0) = 1 \qquad \text{zu} \qquad K = \frac{\Phi^+(t)}{t}\bigg|_{t=0} = (\Phi^+)'(0) \ .$$

Beispiel: $M/M/1$

$$A = M_\lambda : \quad \tilde{A}(t) = \frac{\lambda}{\lambda + t}, \quad B = M_\mu : \quad \tilde{B}(t) = \frac{\mu}{\mu + t}$$

$$\begin{aligned}
\frac{\Psi^+(t)}{\Psi^-(t)} &= \tilde{A}(-t)\tilde{B}(t) - 1 = \frac{\lambda\mu}{(\lambda - t)(\mu + t)} - 1 = \\
&= \frac{t(\mu - \lambda + t)}{(\lambda - t)(\mu + t)}. \\
\Psi^+(t) &= \frac{t(\mu - \lambda + t)}{(t + \mu} \\
\Psi^-(t) &= (\lambda - t) \\
\Phi(z) &= \frac{\Psi^{+'}(0)}{\Psi^+(z)} = \frac{(\mu - \lambda)(\mu + t)}{\mu t(\mu - \lambda + t)} = \frac{1}{t} - \frac{\lambda}{\mu(\mu - \lambda + t)} \\
\Psi^{+'}(0) &= \frac{\Psi^+(t)}{t}\Big|_{t=0} = \frac{\mu - \lambda}{\mu} \\
F(x) &= 1 - \rho e^{-(\mu - \lambda)x} \quad \text{für} \quad x \geq 0
\end{aligned}$$

also die Verteilung der Wartezeit aus dem ersten Kapitel.

Kapitel 4

Abschätzungen und Näherungen

4.1 Abschätzungen

Die Ergebnisse des vorigen Abschnittes sind deswegen etwas unbefriedigend, weil die angestrebte Zerlegung nur in Spezialfällen möglich ist. Im allgemeinen Fall müssen wir uns darauf beschränken, Abschätzungen und näherungsweise Lösungen zu finden.

Wir gehen wieder von unserer Rekursionsgleichung aus:

$$w_{n+1} = (w_n + u_n)_+$$

Wir führen eine neue Zufallsvariable y_n ein, die den entsprechenden Negativteil darstellt:

$$y_n = (w_n + u_n)_- \qquad ((x)_- := (-x)_+) \ .$$

Damit erhalten wir

$$w_{n+1} - y_n = w_n + u_n \ .$$

Das gibt für die Erwartungswerte:

$$\mathbb{E}(w_{n+1}) - \mathbb{E}(y_n) = \mathbb{E}(w_n) + \mathbb{E}(u_n) \ .$$

Für $n \to \infty$ ergibt sich

$$-\mathbb{E}(y) = \mathbb{E}(u) = \mathbb{E}(x) - \mathbb{E}(t) \qquad \text{oder} \qquad \mathbb{E}(y) = \mathbb{E}(t) - \mathbb{E}(x) \ .$$

Leider haben wir keine Beziehung für die Wartezeit (Anmerkung: in den letzten Gleichungen steht nur eine Beziehung für die Verteilungen).

Als nächstes quadrieren wir die Ausgangsgleichung

$$w_{n+1}^2 + y_n^2 - 2w_{n+1}y_n = w_n^2 + 2w_n u_n + u_n^2 \ .$$

Wegen $(x)_+(x)_- = 0$ ist $w_{n+1}y_n = 0$, also

$$w_{n+1}^2 + y_n^2 = w_n^2 + 2w_n u_n + u_n^2 \ .$$

Wenn wir wieder die Erwartungswerte berchnen und $n \to \infty$ gehen lassen, so ergibt sich

$$\begin{aligned} \mathbb{E}(w^2) + \mathbb{E}(y^2) &= \mathbb{E}(w^2) + 2\mathbb{E}(wu) + \mathbb{E}(u^2) = \\ &= \mathbb{E}(w^2) + 2\mathbb{E}(w)\mathbb{E}(u) + \mathbb{E}(u^2) \end{aligned}$$

(w_n und u_n sind ja unabhängig).

Schließlich haben wir

$$\mathbb{E}(w) = \frac{\mathbb{E}(u^2) - \mathbb{E}(y^2)}{-2\mathbb{E}(u)} \qquad [\mathbb{E}(u) \text{ ist ja negativ}] \ .$$

Wir erhalten eine obere Abschätzung, wenn wir $\mathbb{E}(y^2)$ nach unten abschätzen. Dazu verhilft uns die Ungleichung

$$\mathbb{E}(y^2) \geq (\mathbb{E}(y))^2 = (\mathbb{E}(u))^2 \qquad \text{also}$$

$$\mathbb{E}(w^2) \leq \frac{\mathbf{Var}(u)}{-2\mathbb{E}(u)} = \frac{\mathbf{Var}(x) + \mathbf{Var}(t)}{-2\mathbb{E}(u)} \ .$$

Für eine obere Abschätzung für $\mathbb{E}(y^2)$ beachten wir, daß

$$y = (w+u)_- \leq (u)_- \quad \text{da} \quad w \geq 0 \ .$$

Wegen $u^2 = ((u)_+ - (u)_-)^2 = ((u)_+)^2 + ((u)_-)^2$ erhalten wir

$$\mathbb{E}(w) \geq \frac{(\mathbb{E}((u)_+))^2}{-2\mathbb{E}(u)} \ .$$

Ein weiterer Weg, eine untere Abschätzung zu finden ist folgender:

$$\mathbb{E}(w_{n+1}) = \mathbb{E}(w_n + u_n)_+ = \mathbb{E}[\mathbb{E}((w_n + u_n)_+|w_n)] \ .$$

Wenn wir für die Verteilungsfunktion von u_n

$$C(y) = \mathbb{P}(u_n \leq y)$$

setzen, erhalten wir

$$\mathbb{E}((w_n + u_n)_+|w_n = y) = \int_{-y}^{\infty}(u+z)dC(z) = \int_{-y}^{\infty}(1 - C(z))dz =: g(y) \ .$$

Also

$$\mathbb{E}(w_{n+1}) = \mathbb{E}(g(w_n))$$

g ist konvex (g' ist monoton), also können wir die JENSEN'sche Ungleichung anwenden:

$$\mathbb{E}(g(w_n)) \geq g(\mathbb{E}(w_n)) \ .$$

Für $n \to \infty$ ergibt sich

$$\mathbb{E}(w) \geq g(\mathbb{E}(w)) \ .$$

Wir betrachten die Gleichung

$$g(y) = y \ .$$

Die Funktion $y - g(y)$ hat die Ableitung $G(-y) \geq 0$, ist also monoton.
Weiters ist $g(0) = \mathbb{E}((u_n)_+) > 0$ falls $W(u_n > 0) > 0$ ist (andernfalls ist $w = 0$).
Für $n \to \infty$ ist $g(y) \to \mathbb{E}(u) < 0$, es gibt also eine eindeutig bestimmte Lösung y_0, für die $g(y_0) = y_o$, und

$$\mathbb{E}(w) \geq g(y_o) \ .$$

4.2 Näherungen

Ähnlich wie die Abschätzungen des vorigen Kapitels sollen uns die Näherungen dieses Abschnittes dazu dienen, ungefähre Aussagen über das qualitative Verhalten einer Warteschlange zu treffen. Eine Möglichkeit besteht darin, die auftretenden Verteilungen durch solche zu ersetzen, für die wir exakte Ergebnisse kennen. Dazu können wir etwa folgende Klassen von Verteilungen verwenden:

1. **Verteilungen mit rationaler Laplacetransformation**

   Man kann jede Verteilung durch Verteilungen mit rationalen Laplacetransformationen annähern. Für diese Verteilungen kann
   man die Spektralzerlegung für $G/G/1$ „leicht“ durchführen:
   man findet die Nullstellen von Zähler und Nenner und ordnet sie, je nachdem in welcher Halbebene sie liegen, entweder der Funktion Ψ^+ oder Ψ^-zu.

2. **Diskrete Verteilungen**

   Ähnlich wie unter 1. kann man jede Verteilung durch eine diskrete Verteilung annähern. Das folgende Beispiel zeigt, wie die diskreten Verteilungen behandelt werden können:
   Es sei

$$\begin{aligned} \mathbb{P}(t_n = 1) = a, \quad \mathbb{P}(t_n = 2) = 1 - a \\ \mathbb{P}(x_n = 1) = b, \quad \mathbb{P}(x_n = 2) = 1 - b \end{aligned}$$

   $[b > a]$.

   Für $u_n = x_n - t_{n+1}$ ergibt sich:

$$\begin{aligned} \mathbb{P}(u_n = -1) &= b(1-a) \\ \mathbb{P}(u_n = 1) &= a(1-b) \\ \mathbb{P}(u_n = 0) &= ab + (1-a)(1-b) \ . \end{aligned}$$

   Für die stationäre Verteilung der Wartezeit w ergibt sich die Rekursion

$$\begin{aligned} p_k &= a(1-b)p_{k-1} + (ab + (1-a)(1-b))p_k + b(1-a)p_{k+1} \\ p_0 &= (a(1-b) + ab - (1-a)(1-b))p_0 + b(1-a)p_1 \ . \end{aligned}$$

   Wir erhalten

$$\begin{aligned} p_k &= p_0 \left(\frac{a(1-b)}{b(1-a)} \right)^k \\ p_0 &= 1 - \frac{a(1-b)}{b(1-a)} = \frac{b-a}{b(1-a)} \ . \end{aligned}$$

   Falls wir mehr als zwei mögliche Werte für x bzw. t haben, müssen wir eine Rekursion höherer Ordnung lösen; dazu sind bekanntlich die

Nullstellen des charakteristischen Polynoms zu bestimmen. Auch hier, ebenso wie im im vorigen Abschnitt, reduziert sich also das Problem auf die Lösung einer algebraischen Gleichung. Diese Lösung ist für hohe Polynomgrade nur numerisch möglich. Dies und die Tatsache, daß man nicht genau weiß, wie eine „gute“ Näherung zu wählen ist, reduziert die Brauchbarkeit dieser beiden Näherungen.

3. **Approximation für starke Auslastung („Heavy traffic approximation“)**

   Wir betrachten den Fall $\rho \approx 1$.
   Ausgangspunkt unserer Betrachtungen ist die Spektralzerlegung der $G/G/1$:

$$\frac{\Psi^+(s)}{\Psi^-(s)} = \tilde{A}(-s)\tilde{B}(s) - 1 \ .$$

   Für $s \to 0$ erhalten wir daraus die Entwicklung

$$\begin{aligned}
\frac{\Psi^+(s)}{\Psi^-(s)} &= (\mathbb{E}(t) - \mathbb{E}(x))s + \frac{s^2}{2}\mathbb{E}((x-t)^2) + o(s^2) = \\
&= (1-\rho)s\mathbb{E}(t) + \frac{s^2}{2}\mathbb{E}((x-t)^2) + o(s^2) = \\
&= (1-\rho)s\mathbb{E}(t) + \frac{s^2}{2}(\mathbf{Var}(x) + \mathbf{Var}(t) + \\
&+ (1-\rho)^2(\mathbb{E}(t))^2) + o(s^2) \ .
\end{aligned}$$

   Für $\rho \approx 1$ ist $(1-\rho)^2(\mathbb{E}(t))^2$ zu vernachlässigen, also

$$\begin{aligned}
\frac{\Psi^+(s)}{\Psi^-(s)} &\approx (1-\rho)s\mathbb{E}(t) + \frac{s^2}{2}(\mathbf{Var}(x) + \mathbf{Var}(t)) + o(s^2) \approx \\
&\approx s(s - s_0)\frac{\mathbf{Var}(x) + \mathbf{Var}(t)}{2} \ .
\end{aligned}$$

   $\Psi^+(s)$ ist in der Nähe von 0 stetig, also haben wir

$$\Psi^+(s) \approx Cs(s - s_o)$$

   mit

$$s_0 = -\frac{2(1-\rho)\mathbb{E}(t)}{\mathbf{Var}(x) + \mathbf{Var}(t)}$$

und

$$C = \Psi^-(0)\frac{\mathbf{Var}(x) + \mathbf{Var}(t)}{2} \ .$$

Wir erhalten daraus

$$\Phi(s) \approx -\frac{s_0}{s(s - s_0)} = \frac{1}{s} - \frac{1}{s - s_0} \ .$$

Also ergibt sich für die Verteilungsfunktion der Wartezeit

$$F(y) \approx 1 - e^{-y\frac{2\mathbf{E}(t)(1-\rho)}{\mathbf{Var}(x)+\mathbf{Var}(t)}} \ .$$

Die Wartezeit ist also näherungsweise exponentialverteilt mit Mittel

$$\mathbb{E}(w) = \frac{\mathbf{Var}(x) + \mathbf{Var}(t)}{2\mathbb{E}(t)(1 - \rho)} \ .$$

Dieses Ergebnis kann man als „zentralen Grenzwertsatz“ für Warteschlangen betrachten.

Das Mittel dieser Exponentialverteilung haben wir bereits als obere Abschätzung für die mittlere Wartezeit erhalten.

4. **Die Flussapproximation**

   Diese Näherung geht von einer einfachen Idee aus:
   Wir ersetzen die Ankünfte und Bedienvorgänge durch konstante Ströme von λ bzw. μ Kunden pro Zeiteinheit. In unserem Standardmodell ($\lambda < \mu$) ergibt sich aus dieser Näherung natürlich, daß die Schlange stets leer ist, was offensichtlich nicht sehr brauchbar ist.

   Für zwei Fälle gibt diese Approximation aber doch interessante Resultate:

   (a) Falls $\mu < \lambda$ ist, wächst die Anzahl der Kunden im System um $(\lambda - \mu)$ Kunden pro Zeiteinheit.

   (b) Falls $\lambda < \mu$ ist, und falls die Anzahl der Kunden groß ist, kann man die Zeit, bis das System wieder leer ist, mit Hilfe dieser Näherung berechnen.

5. **Die Diffusionsnäherung**

   Dies ist wie die vorige Näherung eine Approximation durch einen kontinuierlichen Prozeß. Diesmal wird auch die Varianz betrachtet.

Es sei $N_a(u)$ die Anzahl der Kunden, die bis zur Zeit t ankommen. Es gilt die Beziehung

$$N_a(u) \geq n \Leftrightarrow T_n \geq u$$

T_n ist, wie üblich, die Ankunftszeit des n-ten Kunden.

Aus dem Gesetz der großen Zahlen folgt:

$$T_n \approx n\mathbb{E}(t) \ .$$

Das impliziert

$$N_a(u) \approx \lambda u \ .$$

Der zentrale Grenzwertsatz gibt uns

$$\mathbb{P}(T_n \leq n\mathbb{E}(t) + z\sqrt{n\mathbf{Var}(t)}) = \Phi(z) \ .$$

Daraus ergibt sich für großes n

$$\begin{aligned}
&\mathbb{P}(N_a \geq \lambda u + y\sqrt{u}) = \\
&= \mathbb{P}(T_{\lambda u + y\sqrt{u}} \leq u) = \\
&= \mathbb{P}(T_{\lambda u + y\sqrt{u}} \leq \mathbb{E}(t)(\lambda u + y\sqrt{u}) - y\mathbb{E}(t)\sqrt{u}) = \\
&= \mathbb{P}(T_{\lambda u + y\sqrt{u}} \leq \mathbb{E}(t)(\lambda u + y\sqrt{u}) - \\
&\quad -y\sqrt{(\mathbb{E}(t))^3(\lambda u + y\sqrt{u})}) = \\
&= \mathbb{P}(T_{\lambda u + y\sqrt{u}} \leq \mathbb{E}(t)(\lambda u + y\sqrt{u}) - \\
&\quad -\frac{y}{\sqrt{\lambda^3\mathbf{Var}(t)}}\sqrt{\mathbf{Var}(t)(\lambda u + y\sqrt{u})}) \\
&= 1 - \Phi(\frac{y}{\sqrt{\lambda^3\mathbf{Var}(t)}}) \ .
\end{aligned}$$

$N_a(u)$ ist also näherungsweise normalverteilt mit Mittel $u\lambda$ und Varianz $u\lambda^3\mathbf{Var}(t)$. Genauso erhält man für die Anzahl $N_b(u)$ der Kunden, die während der Zeit u bedient werden (wenn die Schlange nicht leer ist) eine näherungsweise Normalverteilung mit Mittel μu und Varianz $\mu^3 u\mathbf{Var}(x)$. Wir nehmen jetzt an, daß diese Werte durch kontinuierliche Beiträge zustande kommen, d.h. die „Anzahl“ der Ankünfte (bzw. Bedienvorgänge) in der kurzen Zeit Δu soll normalverteilt mit Mittel

$\lambda\Delta u$ (bzw. $\mu\Delta u$) und Varianz $\lambda^3\Delta u\mathbf{Var}(t)$ (bzw. $\mu^3\Delta u\mathbf{Var}(x)$) sein. Die Anzahl der Ankünfte bzw. Bedienvorgänge über disjunkten Intervallen sollen natürlich unabhängig sein. Die Änderung der Anzahl der Kunden im System während der Zeit Δu ist dann normalverteilt mit Mittel $\Delta u(\lambda - \mu)$ und Varianz $\Delta u\sigma^2$ mit $\sigma^2 = \lambda^3\mathbf{Var}(t) + \mu^3\mathbf{Var}(x)$. (Die letzte Beziehung gilt natürlich nur, wenn die Anzahl der Kunden im System > 0 ist).

Es sei nun

$$F(x,u) = \mathbb{P}(N(u) \leq x) \ .$$

Wir stellen eine Gleichung für $F(x, u + \Delta u)$ auf, dabei sei $X(\Delta u)$ die Änderung der Kunden während Δu:

$$\begin{aligned}
F(x,u+\Delta u) &= \mathbb{P}(N(u+\Delta u) \leq x) = \\
&= \mathbb{P}(N(u) + X(\Delta u) \leq x) = \\
&= \mathbb{P}(N(u) \leq x - X(\Delta u)) = \\
&= \mathbb{E}(F(x - X(\Delta u), u)) = \\
&= \mathbb{E}(F(x,u) - F_x(x,u)X(\Delta u) + \\
&\quad + \frac{1}{2}F_{xx}(x,u)X(\Delta u)^2 + o(\Delta u)) = \\
&= F(x,u) - F_x(x,u)\mathbb{E}(X(\Delta u)) + \\
&\quad + \frac{1}{2}F_{xx}(x,u)(\mathbb{E}(X(\Delta u)^2)) + o(\Delta u) = \\
&= F(x,u) - F_x(x,u)\Delta u(\lambda - \mu) + \\
&\quad + \frac{1}{2}F_{xx}(x,u)\sigma^2\Delta u + o(\Delta u) \ .
\end{aligned}$$

Wenn man $F(x,u)$ nach links bringt, durch Δu dividiert, und $\Delta u \to 0$ gehen läßt, ergibt sich

$$F_u(x,u) = (\mu - \lambda)F_x(x,u) + \frac{1}{2}\sigma^2 F_{xx}(x,u) \qquad (x \geq 0)$$

$$\begin{aligned}
F(x,0) &= 1 \qquad (x \geq 0) \\
F(x,0) &= 0 \qquad (x < 0) \\
F(0,u) &= 0 \qquad (u > 0) \ .
\end{aligned}$$

Die Anfangsbedingung ergibt sich aus der Forderung, daß das System zur Zeit 0 leer sein soll, und die Randbedingung daraus, daß die Anzahl der Kunden nicht negativ sein darf.

Man kann sehr leicht die Lösung der Gleichung ohne Randbedingung finden, indem man die Laplace-Transformation anwendet:

$$G(z,u) = \int_{-\infty}^{\infty} e^{-xz} F(x,u) dx \ .$$

Wir erhalten nach einigen partiellen Integrationen:

$$G_u(z,u) = (\mu - \lambda) z G(z,u) + \frac{1}{2}\lambda^2 z^2 G(z,u) \ ,$$

also

$$G(z,u) = G(z,0) e^{\frac{1}{2}\sigma^2 z^2 + (\mu-\lambda)z}$$

und, da $G(z,0) = \frac{1}{z}$

$$G(z,u) = \frac{1}{z} e^{\frac{1}{2}\sigma^2 u z^2 + (\mu-\lambda)zu} \ .$$

Die Inversion der Laplace-Transformation liefert

$$F(x,u) = \Phi\left(\frac{x + (\mu-\lambda)u}{\sqrt{\sigma^2 u}}\right) \ .$$

Um die Gleichung mit der Randbedingung zu lösen, suchen wir zuerst die stationäre Verteilung. Diese ergibt sich aus der obigen Gleichung, wenn $F(x,u)$ nicht von u abhängt. Dann ist natürlich die Ableitung nach $u = 0$, und wir erhalten:

$$F_0'(x)(\mu - \lambda) + \frac{1}{2}\sigma^2 F''(x) = 0 \ .$$

Diese Gleichung hat die allgemeine Lösung

$$F(x) = C_1 + C_2 e^{\frac{2(\lambda-\mu)}{\sigma^2}x} \ .$$

Da $F(0) = 0$ und $F(\infty) = 1$ sein soll, erhalten wir

$$F(x) = 1 - e^{\frac{2(\lambda-\mu)}{\sigma^2}x} \ .$$

Es ergibt sich also wieder eine Exponentialverteilung für die Wartezeit. Für $\rho \to 1$ stimmt diese Verteilung in erster Näherung mit der aus Abschnitt 3. überein. Mit etwas mehr Arbeit kann man die folgende Lösung der partiellen Differentialgleichung erhalten:

$$F(x,u) = \Phi\left(\frac{x + u(\mu-\lambda)}{\sqrt{\sigma^2 u}}\right) - e^{\frac{2(\mu-\lambda)}{\sigma^2}x} \Phi\left(\frac{-x + u(\mu-\lambda)}{\sigma^2}\right) \ .$$

Kapitel 5

Time-Sharing

Wir wollen jetzt unsere Kenntnisse auf eine Analyse von Fragen anwenden, die bei Time-sharing Anwendungen auftreten.
Wir betrachten den einfachsten Fall, daß nur eine CPU vorhanden ist, die „gleichzeitig“ mehrere Programme bearbeiten muß. Dazu wird jeweils eines der wartenden Programme ausgewählt, in den Hauptspeicher geladen, eine kurze Zeit (die sog. „Zeitscheibe“) gerechnet, aus dem Hauptspeicher entfernt, und das Spiel mit dem nächsten Programm fortgesetzt. Jedes Programm braucht eine bestimmte Rechenzeit x, und sobald diese Zeit (in Form von einzelnen Zeitscheiben) abgearbeitet ist, verläßt das Programm das System. Da wir keine a-priori Information über die Rechenzeit eines Programmes voraussetzen, können wir die Jobs nur aufgrund der schon verbrauchten Rechenzeit klassifizieren, und die Auswahl des nächsten Programms nach dieser verbrauchten Rechenzeit treffen. Dabei können wir verschiedene Ziele verfolgen:

1. kurze Programme sollen möglichst schnell erledigt werden. Dadurch wird die Anzahl der Programme im System klein gehalten, was den Verwaltungsaufwand reduziert; außerdem ist es psychologisch ungünstig, wenn ein Kunde auf ein 2-Sekunden-Programm eine Stunde warten muß.

2. eine möglichst „gerechte“ Verteilung wäre eine, bei der die Zeit, die ein Job im System verbringt, proportional zur Rechenzeit ist; nur dann ist es nicht möglich, durch Aufteilen eines langen Programmes in mehrere

kürzere bzw. durch Zusammenfassen mehrere kürzere Programme einen Zeitgewinn zu erzielen.

Wir machen folgende Annahmen:

1. Die Ankünfte erfolgen nach einem Poissonprozeß mit Rate λ, die Rechenzeiten sind unabhängig mit Verteilungsfunktion B (wir haben also eine $M/G/1$-Situation) mit Dichte b.
2. Die Zeitscheibe nehmen wir als infinitesimal an; weiters nehmen wir an, daß wir die Zeit zum Austauschen vernachlässigen können.
3. Wir betrachten nur die stationären Verteilungen.

$N(u)$ sei die mittlere Anzahl von Jobs, deren verbrauchte Rechenzeit $\leq u$ ist. Dazu möge eine Dichte $n(u)$ existieren, sodaß

$$N(u) = \int_0^u n(s)ds$$

ist.

$T(u)$ sei die Zeit, die im Durchschnitt vergeht, bis ein Job u Sekunden Rechenzeit bekommt.
$W(u)$ sei die Wartezeit eines Jobs mit u Sekunden Rechenzeit, also

$$W(u) = T(u) - u \ .$$

Wir betrachten die Jobs, die schon zwischen u und $u+\Delta u$ Sekunden gerechnet haben, als eine eigene Warteschlange. Hier kommen alle Jobs durch, deren Rechenzeit $\geq u$ ist. Die Ankunftsrate in dieser Schlange ist also $\lambda(1-B(u))$. Die mittlere Aufenthaltsdauer ist $T(u+\Delta u)-T(u)$, und die mittlere Anzahl von Jobs in dieser Schlange ist $\approx n(u)\Delta u$.
Mithilfe des Satzes von Little ergibt sich die Beziehung

$$n(u) = \lambda(1 - B(u))\frac{dT(u)}{du} \ .$$

Wir betrachten die folgende Strategien:

1. **FCFS** („Batch“)

2. **LCFS** (prä-emptiv): ein Job, der das System betritt, wird sofort bearbeitet (ein eventuell laufender Job wird dazu unterbrochen); wenn ein Job fertig ist, wird der zuletzt unterbrochene weiterbearbeitet.

3. **Round Robin (RR)**: alle Jobs, die im System sind, werden der Reihe nach bearbeitet (abwechselnd).

4. Es wird jeweils der Job bearbeitet, der am wenigsten Rechenzeit verbraucht hat.

Es sollte Strategie 4 kurze Jobs am meisten bevorzugen, 1 am wenigsten, 2 und 3 sollten dazwischen liegen.

1. Kennen wir von früher; hier ist die Wartezeit $W(u)$ konstant, und zwar ist
$$W(u) = \frac{\lambda \mathbb{E}(x^2)}{2(1-\rho)}$$
und
$$T(u) = u + \frac{\lambda \mathbb{E}(x^2)}{2(1-\rho)} \ .$$

2. Hier ist $T(u)$ leicht zu bestimmen. Denn $T(u)$ ist gleich der Rechenzeit u plus der Summe der Rechenzeiten aller Programme, die während $T(u)$ ankommen. Während $T(u)$ kommen $\lambda T(u)$ Programme an, jedes bringt im Schnitt $\mathbb{E}(x)$ Sekunden Rechenzeit mit, also gilt
$$T(u) = u + \lambda T(u)\mathbb{E}(x) = u + \rho T(u) \ ,$$
also
$$T(u) = \frac{u}{1-\rho} \ .$$
Wir haben also ein „gerechtes“ Verfahren gefunden.

3. Wenn sich N Programme im System befinden, bekommt ein bestimmtes Programm $\frac{1}{N}$ der gesamten Rechenzeit.
Daher ist $dT(u) = N du$. Da nur gerechnet wird, wenn das System nicht leer ist, ergibt sich:
$$T(u) = u\mathbb{E}(N \mid N \neq 0) = Cu,$$

also wieder ein „gerechtes“ System. Um C zu bestimmen, betrachten wir den Fall $u \to \infty$. Wenn u groß ist, werden die meisten Jobs, die während $T(u)$ ankommen, auch noch während $T(u)$ das System verlassen. Für großes u ist also das Verhalten ähnlich wie im vorigen Fall, und wir erhalten wieder

$$T(u) = \frac{u}{1-\rho} \ .$$

4. Wenn wir ein Programm betrachten, das genau u Sekunden Rechenzeit benötigt, dann sehen wir, daß für $T(u)$ von der Rechenzeit aller anderen Programme nur der Teil von Bedeutung ist, der kleiner als u ist. Aus der Sicht dieses Programmes können wir also x durch $(x \wedge u)$ ersetzen, und die Verteilungsfunktion B durch:

$$B_u(y) = \begin{cases} B(y) & y < u \\ 1 & y \geq u \end{cases} \ .$$

$W(u)$ setzt sich jetzt zusammen aus der restlichen Rechenzeit aller Programme, die vor unserem Programm angekommen sind, plus der Summe der Rechenzeiten von allen Programmen, die während $T(u)$ ankommen. Der erste Teil ist im Mittel gleich der Wartezeit in $M_\lambda/B_u/1$, also gleich

$$W_u = \frac{\lambda \mathbb{E}((x \wedge u)^2)}{2(1-\rho_u)}$$

mit

$$\rho_u = \lambda \mathbb{E}(x \wedge u) \ .$$

Für den zweiten Teil ergibt sich

$$\lambda T(u) \mathbb{E}(x \wedge u) = T(u)\rho_u \ .$$

Wir bekommen die Gleichung

$$T(u) = u + W_u + \rho_u T(u) \ ,$$

also

$$T(u) = \frac{u + W_u}{1-\rho_u} \ .$$

Für $u \to 0$ ergibt sich

$$T(u) \approx u \ ,$$

für $u \to \infty$

$$T(u) \approx \frac{u}{1-\rho} \ .$$

Kapitel 6

Prioritäten

Wir betrachten den Fall, daß es mehrere Klassen von Kunden gibt, die von unserem System unterschiedlich behandelt werden. Genauer gesagt soll es $p > 0$ Klassen von Kunden geben. Die Kunden aus Klasse i kommen nach einem Poissonprozeß mit Rate λ_i an und benötigen eine Bedienzeit mit Verteilungsfunktion B_i (wir betrachten also wieder eine $M/G/1$-Situation). Weiters sei

$$\begin{aligned} \lambda &= \sum_{i=1}^{p} \lambda_i \\ B(y) &= \frac{1}{\lambda} \sum_{i=1}^{p} \lambda_i B_i(y) \\ \rho_i &= \lambda_i \int y dB_i(y) \\ \rho &= \lambda \int y dB(y) \ . \end{aligned}$$

Es gibt jetzt eine ganze Reihe von Möglichkeiten, Disziplinen zu definieren, die eine Klasse gegebüber anderen bevorzugen. Wir sprechen von unterschiedlichen **Prioritäten**.

Die Disziplinen, die wir untersuchen, sollen folgende Eigenschaften haben:

1. **Nicht-prä-emptiv**: Ein einmal begonnener Bedienvorgang wird ohne Unterbrechung zu Ende geführt.

2. **Arbeitserhaltend**: Niemand, der wartet, wird weggeschickt, ohne bedient zu worden zu sein.

Weiters soll immer, wenn das System nicht leer ist, bedient werden.

6.1 Ein Erhaltungssatz

U_t sei die unverrichtete Zeit im System, d.h. die Zeit, die benötigt wird, um alle anwesenden Kunden fertig zu bedienen. Offensichtlich ist die Verteilung von U_t unabhängig von der Disziplin:
U_t wächst mit jeder Ankunft eines Kunden um seine Bedienzeit an, und fällt pro Sekunde um 1 Sekunde, solange $U_t > 0$ ist. Die stationäre Verteilung von U_t entspricht der Verteilung der Wartezeit eines zufällig ankommenden Kunden bei der FCFS-Disziplin.
Insbesondere ist

$$\mathbb{E}(U) = \frac{\lambda \mathbb{E}(x^2)}{2(1-\rho)}$$

wobei x nach der Funktion B verteilt ist.

Wir berechnen jetzt $\mathbb{E}(U)$ auf eine andere Art. Dazu bezeichnen wir mit W_i, $i = 1, \ldots, p$, die mittlere Wartezeit eines Kunden mit Priorität i, und mit N_i die Anzahl der Kunden aus der i-ten Prioritätsgruppe in der Warteschlange.

$\mathbb{E}(U)$ setzt sich jetzt aus folgenden Beiträgen zusammen:

1. W_0: die mittlere restliche Bedienzeit für den Kunden, der gerade bedient wird.

2. Die Summe der mittleren Bedienzeiten für alle Kunden, die sich in der Warteschlange befinden.

Um W_0 zu bestimmen, stellen wir fest, daß W_0 gleich der Zeit ist, die ein zufällig ankommender Kunde warten muß, bis der Kunde fertig ist, der gerade bedient wird. Mit Wahrscheinlichkeit $(1-\rho)$ findet der ankommende Kunde das System leer vor. Falls der Server besetzt ist, kann man die Verteilung der restlichen Bedienzeit folgendermaßen bestimmen:

Wir betrachten eine große Anzahl n von unabhängigen Variablen mit Verteilung B. Ihre Summe ist nach dem Gesetz der großen Zahlen $\approx n\mathbb{E}(x)$. Wenn wir in dem Intervall der Länge $n\mathbb{E}(x)$ einen Punkt zufällig wählen, ist die Chance, daß wir bis zum Ende des Teilintervalls, in das der zufällig gewählte Punkt fällt, einen Abstand zwischen u und $u + \Delta u$ haben, gleich $\frac{\Delta u}{n\mathbb{E}(x)}$ mal der Anzahl der Intervalle mit Länge $> u$, also

$$\approx \frac{\Delta u}{n\mathbb{E}(x)} n(1 - B(u)) \ .$$

Für $n \to \infty$ ergibt sich für die Verteilung der restlichen Wartezeit die Dichte

$$f(u) = \frac{1 - B(u)}{\mathbb{E}(x)} \ .$$

Schließlich ist

$$W_0 = \rho \int_0^\infty u f(u) du = \frac{\rho\mathbb{E}(x^2)}{2\mathbb{E}(x)} = \frac{\lambda\mathbb{E}(x^2)}{2} \ .$$

Die Summe der Bedienzeiten der Kunden in der Warteschlange ergibt sich natürlich aus der Summe der gesamten Bedienzeiten der Kunden in den einzelnen Gruppen. Es sind im Schnitt N_i Kunden aus der i-ten Gruppe anwesend. Für jeden wird im Schnitt eine Bedienzeit $\mathbb{E}(x_i)$ (x_i soll eine ZV mit Verteilung B_i sein) benötigt.
Damit gilt

$$\mathbb{E}(U) = W_0 + \sum_{i=1}^{p} \mathbb{E}(x_i) N_i = \frac{W_0}{1 - \rho} \ .$$

Nach Little gilt $N_i = \lambda_i W_i$, also schließlich

$$\sum_{i=1}^{p} \rho_i W_i = \frac{\rho W_0}{1 - \rho} = \frac{\lambda\rho\mathbb{E}(x^2)}{2(1 - \rho)} \ .$$

Dieses Ergebnis zeigt, daß wir eine Gruppe nur bevorzugen können, indem eine andere Gruppe größere Wartezeiten in Kauf nehmen muß.

6.2 Eine Methode zur Bestimmung der Wartezeit

Wir betrachten einen Kunden aus der Gruppe i, der das System betritt:
N_{ij} ... mittlere Anzahl der Kunden aus Gruppe j, die unser Kunde im Sys-

tem antrifft, und die vor ihm bedient werden (ausgenommen der Kunde, der eventuell gerade bedient werden, wenn unser Kunde ankommt).
M_{ij} ... mittlere Anzahl der Kunden aus Gruppe j, die während der Wartezeit unseres Kunden ankommen und vor ihm bedient werden.
Damit gilt

$$W_i = W_0 + \sum_{j=1}^{p} (N_{ij} + M_{ij}) \mathbb{E}(x_j) \ .$$

Wir verwenden diesen Zugang für die einfachste Disziplin:
Jeder Kunde aus Gruppe i wird vor allen Kunden aus Gruppe $i-1$ bedient, und innerhalb einer Gruppe wird nach $FCFS$ gearbeitet.
Dann ist

$$\begin{aligned} N_{ij} &= 0 \qquad j < i \\ M_{ij} &= 0 \qquad j \leq i \ . \end{aligned}$$

Für $j \geq i$ ist

$$N_{ij} = N_j = \lambda_j W_j \ .$$

Für $j > i$ ist M_{ij} die Anzahl der Kunden aus Gruppe j, die im Mittel während $W_i = \lambda_j W_i$ ankommen.

Wir erhalten

$$\begin{aligned} W_i &= W_0 + \sum_{j=i}^{p} \rho_j W_j + \sum_{j=i+1}^{p} \rho_j W_i = \\ &= W_0 + W_i \sum_{j=i}^{p} \rho_j + \sum_{j=i+1}^{p} \rho_j W_j \end{aligned}$$

oder

$$W_i (1 - \sum_{j=i}^{p} \rho_j) = W_0 + \sum_{j=i+1}^{p} \rho_j W_j \ .$$

Wir schreiben

$$\sigma_j = \sum_{j=i}^{p} \rho_j$$

und erhalten

$$W_{i-1}(1 - \sigma_{i-1}) = W_i (1 - \sigma_i) + \rho_i W_i = W_i (1 - \sigma_{i+1}) \ ,$$

und schließlich

$$W_i = \frac{W_0}{(1 - \sigma_i)(1 - \sigma_{i+1})} \ .$$

Anhang A

Transformationen

Für unsere Untersuchungen benötigen wir die folgenden Transformationen:

1. Die erzeugende Funktion oder z - Transformierte: Falls p_n, $n \geq 0$ eine diskrete Verteilung ist, nennen wir

   $$P^*(z) = \sum p_n z^n$$

   die erzeugende Funktion von (p_n). Falls X die Verteilung (p_n) hat, so gilt

   $$P^*(z) = \mathbb{E}(z^X) \ .$$

   $P^*(z)$ existiert jedenfalls für $|z| < 1$. Bekanntlich kann man aus P^* eindeutig (p_n) bestimmen:

   $$p_n = \frac{1}{n!} P^{*^n}(0) \ .$$

2. Die Laplace - Transformierte: Falls $f(x)$, $x \geq 0$ eine Dichtefunktion ist, d.h.

   $$f \geq 0 \qquad \text{und} \qquad \int_0^\infty f(x)dx = 1 \ ,$$

   heißt

   $$\hat{F}(t) = \int_0^\infty e^{-xt} f(x)dx$$

die Laplace - Transformierte von f. Dieses Integral ist für $t \geq 0$ endlich, und es gibt auch hier eine eindeutige Beziehung zwischen $\hat{F}$ und f. Falls X mit der Dichte f verteilt ist, ist

$$\hat{F}(t) = \mathbb{E}(e^{-Xt}) \ .$$

Diese Beziehung kann man auch verwenden, um die Laplace - Transformierte für nicht stetige Verteilungen zu definieren.

Es bestehen folgende Eigenschaften der Transformationen:

1. $P^*(z)$ ist regulär für $|z| \leq 1$.
2. $\hat{F}(z)$ ist regulär für $\Re(z) \geq 0$. Falls X eine Verteilung (p_n) hat, ist
$$\mathbb{E}(X) = (P^*)'(1) \ .$$
Falls X Dichte f hat, ist
$$\mathbb{E}(X) = -\hat{F}'(0) \ .$$
3. Falls X, T unabhängig sind, ist die Transformierte der Summe das Produkt der Transformierten.
4. Weiters sei N_t ein Poissonprozeß (d.h. eine Folge von Ereignissen, wobei die Zeit zwischen zwei Ereignissen nach M_λ verteilt ist. N_t ist die Anzahl dieser Ereignisse im Intervall $[0, t]$). Für eine zufällige Zeit T wollen wir die Anzahl N_T von Ereignissen in $[0, T]$ bestimmen. Falls $T = t$ ist, ist diese Anzahl Poisson - verteilt mit Parameter λt:
$$\mathbb{P}(N_T = n | T = t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \ ,$$
also ist
$$\mathbb{P}(N_T = n) = \mathbb{E}\left[\frac{(\lambda T)^n}{n!} e^{-\lambda T}\right] \ .$$
Die erzeugende Funktion ist
$$\begin{aligned} \hat{\mathbb{P}}(z) &= \sum \mathbb{P}(N_T = n) z^n = \\ &= \mathbb{E}\left[\sum e^{-\lambda T} \frac{(\lambda z T)^n}{n!}\right] = \\ &= \mathbb{E}(e^{-\lambda(1-z)T}) = \\ &= \tilde{F}(\lambda(1-z)) \ , \end{aligned}$$

falls T mit Dichte f verteilt ist.