

# Warteschlangentheorie

Karl Grill

Institut für Stochastik und Wirtschaftsmathematik  
([karl.grill@tuwien.ac.at](mailto:karl.grill@tuwien.ac.at))

Version 1.2016.1 “Denali”  
©1999–2016 Karl Grill CC-BY-SA

# Vorwort

Dieses Skriptum, das schon etwas in die Jahre (17, wie in alten Schlagern) gekommen ist, bekommt jetzt endlich einige überfällige Korrekturen und Ergänzungen. In der gegenwärtigen Version sind das hauptsächlich Korrekturen der Druckfehler der ersten Version. In der näheren Zukunft wird noch einiges Material, das in der Vorlesung präsentiert wurde, dazukommen. Ich bedanke mich bei Frau Gieraga vom TU-Verlag für den Impuls dafür, dieses Projekt endlich in Angriff zu nehmen.

Was auch schon sehr in die Jahre gekommen ist, ist die Art, wie hier mit den Geschlechtern umgegangen wurde (alle Mann an Bord!). Leider sind die ehrlichen Ansätze zum gender mainstreaming alle ein wenig umständlich (abgesehen von der totalen Vermeidung, aber irgendwie fühlt sich die auch nicht so ehrlich an), deshalb werde ich mich hier durchschummeln, indem ich die vorhandenen Rollen abwechselnd mit Frauen und Männern besetze, wobei im Zweifelsfall die Frauen die besseren Rollen bekommen.

Zu guter Letzt möchte ich noch meinen Dank an Pantelis Christodoulides und Klaus Berger ausdrücken, die die erste Version dieses Skriptums erstellt haben.

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>4</b>
<b>2</b>	<b>Erste Resultate</b>	<b>7</b>
2.1	Eine Rekursion für die Wartezeit . . . . .	7
2.2	Der Satz von Little . . . . .	11
<b>3</b>	<b>Warteschlangensysteme</b>	<b>14</b>
3.1	Die Schlange $M/M/1$ . . . . .	14
3.2	Das System $M/G/1$ . . . . .	16
3.3	Das System $G/M/1$ . . . . .	18
3.4	Das System $G/G/1$ . . . . .	19
<b>4</b>	<b>Abschätzungen und Näherungen</b>	<b>22</b>
4.1	Abschätzungen . . . . .	22
4.2	Näherungen . . . . .	24
<b>5</b>	<b>Der Einfluss der Disziplin</b>	<b>31</b>
5.1	Nichtpräemptive Disziplinen - FCFS und ihre Freundinnen	31
5.2	Präemptive Disziplinen — Time-Sharing . . . . .	32
5.3	Prioritäten . . . . .	36
5.3.1	Ein Erhaltungssatz . . . . .	36
5.3.2	Eine Methode zur Bestimmung der Wartezeit . . . . .	38



# Kapitel 1

## Einleitung

Warteschlangen kennen wir alle aus unserer täglichen Anschauung.



Abbildung 1.1: Eine Warteschlange

Wir einigen uns auf das folgende Modell: Kundinnen kommen zu zufälligen Zeiten  $0 < T_1 < T_2 < \dots < T_n < \dots$  im System an, wobei  $T_n$  die Ankunftszeit der  $n$ -ten Kundin bezeichnet, wobei wir zusätzlich noch  $T_0 = 0$  definieren (unser System beginnt also zur Zeit 0 zu arbeiten, und zwar im leeren Zustand. Zur Vereinfachung der Beschreibung führen wir die *Zwischenankunftszeiten*

$$t_n = T_n - T_{n-1}, \quad n = 1, \dots$$

ein.

Ein oder mehrere Bediener arbeiten die Schlange ab, und für jede Kundin wird eine bestimmte ‘Bedienzeit’ benötigt. Wir bezeichnen die Bedienzeit der  $n$ -ten Kundin. mit  $x_n$ .

Die Reihenfolge der Bedienung der Kundinnen wird durch die sogenannte ‘Disziplin’ der Warteschlange bestimmt. Am gebräuchlichsten (zumindest wenn Menschen involviert sind) ist die Abarbeitung in der Reihenfolge der Ankünfte, FCFS (First Come First Served).

Disziplinen legen fest, ob der Server Pausen machen kann, ob ein einmal begonnener Bedienvorgang unterbrochen werden kann (ist das der Fall, nennen wir die Disziplin präemptiv, sonst nicht-präemptiv), und schließlich, welche Kundin für die nächste Bedienung ausgewählt wird. Neben FCFS ist hier etwa die umgekehrte Reihenfolge der Ankünfte (last come first served, LCFS), Prioritäten (gewisse Kundinnen werden gegenüber anderen bevorzugt) oder eine Zufallsauswahl möglich.

Wir werden meistens annehmen, dass

1. die Bedienzeiten  $x_n$  unabhängig und identisch verteilt sind mit einer Verteilungsfunktion  $B$ ,
2. die Zwischenankunftszeiten  $t_n$  ebenfalls unabhängig und identisch verteilt sind, mit Verteilungsfunktion  $A$ ,
3. und dass die Folgen  $(x_n)$  und  $(t_n)$  auch voneinander unabhängig sind.

Mit diesen Voraussetzungen können wir unsere Warteschlangen mithilfe der *Kurznotation nach Kendall* angeben: wir schreiben:  $A/B/s$ . Für eine Schlange mit  $s$  Servern, Verteilung  $A$  für die Zwischenankunftszeiten, Verteilung  $B$  für die Bedienzeiten und FCFS als Disziplin. Abweichungen von diesen Standardvorgaben werden oft als zusätzliche Angaben daran angehängt (etwa  $A/B/s/LCFS$  oder  $A/B/s/N$  für eine Warteschlange mit begrenztem Warteraum mit  $N$  Plätzen), aber diese sind nicht standardisiert.

Für die Verteilungen gibt es einige spezielle Kürzel:

$M \dots$  Exponentialverteilung ('memoryless').

Dichtefunktion:

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

$E_n \dots$  Erlangverteilung: eine Summe von  $n$  unabhängigen identischen Exponentialverteilungen.

Dichtefunktion:

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

$H \dots$  Hyperexponentielle Verteilung: eine Mischung von Exponentialverteilungen. Wir haben  $n$  Gewichte  $p_1 \dots p_n$ ,  $p_i \geq 0$ , mit  $\sum_{i=1}^n p_i = 1$  und  $n$  Parameterwerte  $\lambda_1 \dots \lambda_n > 0$ .

Damit ist die 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. Diese Eigenschaft charakterisiert die Exponentialverteilung.

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

1.  $w_n$  ... Wartezeit der  $n$ -ten Kundin.
2.  $z_n = w_n + x_n$  ... Zeit, die die  $n$ -te Kundin im System verbringt.
3.  $N_t$  ... Anzahl der Kundinnen, 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  $n \rightarrow \infty$  bzw.  $t \rightarrow \infty$  (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 der  $(n+1)$ -ten Kundin durch die des  $n$ -ten Kunden ausdrücken. Dazu ist

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

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

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

Wir setzen  $u_n = x_n - t_{n+1}$ . Die  $u_n$  sind unter unseren üblichen Unabhängigkeitsannahmen für die Zwischenankunfts- und Bedienzeiten unabhängig und identisch verteilt. Mit dieser Notation erhält die Rekursion die Form

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

*Anmerkung:* die Folge  $(w_n)$  bildet unter diesen Voraussetzungen einen (homogenen) Markovprozess.



Für uns ist der Fall besonders interessant, dass die Folge  $(w_n)$  nicht ins Unermessliche wächst, sondern (in Verteilung/Wahrscheinlichkeit) beschränkt bleibt. Um eine Bedingung dafür zu finden, iterieren wir die Rekursion:

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

Die Folge  $(u_{n-1}, \dots, u_1)$  hat, weil die  $u_n$  unabhängig und identisch verteilt sind, dieselbe Verteilung wie  $(u_1, \dots, u_{n-1})$ . Also ist die Verteilung von  $w_n$  dieselbe wie die von

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

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

$$\tilde{w} = \lim_{n \rightarrow \infty} \tilde{w}_n .$$

Wenn  $\tilde{w}$  endlich ist, dann ist seine Verteilung die Grenzverteilung der Folge  $w_n$ , und diese Verteilung ist eine stationäre Verteilung für die Warteschlange — wenn wir statt der üblichen Annahme  $w_1 = 0$  diese Verteilung geben, dann hat für jedes  $n$   $w_n$  dieselbe Verteilung, und wegen der Markoveigenschaft bildet dann  $(w_n)$  eine stationäre Folge.

Nach dem starken Gesetz der großen Zahlen gilt mit Wahrscheinlichkeit 1

$$(u_1 + \dots + u_n)/n \rightarrow \mathbb{E}(u) .$$

Falls  $\mathbb{E}(u) > 0$ , dann geht  $u_1 + \dots + u_{n-1} \rightarrow \infty$ , also gilt  $\tilde{w} = \infty$ . Falls  $\mathbb{E}u < 0$ , dann geht  $u_1 + \dots + u_{n-1} \rightarrow -\infty$ , also ist für  $n > n_0$   $u_1 + \dots + u_{n-1} < 0$ , was bedeutet dass nur die ersten Glieder in der Definition von  $\tilde{w}_n$  wichtig sind; also ist  $\tilde{w}$  endlich. Falls  $\mathbb{E}u = 0$ , 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

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

Somit ist für jedes  $n$

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

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’  $\lambda = \frac{1}{\mathbb{E}(t)}$  und  $\mu = \frac{1}{\mathbb{E}(x)}$  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. Zusätzlich definieren wir die Auslastung  $\rho = \frac{\lambda}{\mu}$ . Mit diesen Bezeichnungen wird die Bedingung für stationäres Verhalten zu  $\lambda < \mu$  bzw.  $\rho < 1$ .

Nach diesem schönen Erfolg werden wir unbescheiden und wollen noch einen endlichen Erwartungswert für die Grenzverteilung der Wartezeit (in einigen Fällen werden wir diesen Erwartungswert explizit angeben können).

Wir nehmen also an, dass  $\rho < 1$  gilt und speziell, dass die Erwartungswerte von Zwischenankunfts- und Bedienzeit endlich sind. Das garantiert zwar, dass  $\mathbb{E}(w_n) = \mathbb{E}(\tilde{w}_n)$  endlich ist, aber für die Endlichkeit von  $\mathbb{E}(\tilde{w})$  ist etwas mehr nötig, wie sich zeigen wird.

Der Erwartungswert von  $\tilde{w}$  ist genau dann endlich, wenn für beliebiges  $c > 0$  die Reihe

$$\sum_n \mathbb{P}(\tilde{w} > cn)$$

konvergiert. Wir wählen  $a > \mathbb{E}(t) > \mathbb{E}(x)$ . Wir schätzen ab:

$$\begin{aligned} \mathbb{P}(\tilde{w} > cn) &\geq \mathbb{P}(u_1 + \dots + u_n > cn) \geq \\ &\sum_{i=1}^n \mathbb{P}(x_i > (a+c)n, \sum_{1 \leq j \leq n, j \neq i} x_j < an, \sum_{j=1}^n t_j < an) = \\ &n \mathbb{P}(x_1 > (a+c)n, \sum_{2 \leq j \leq n} x_j < an, \sum_{j=1}^n t_j < an) = \\ &n \mathbb{P}(x_1 > (a+c)n) \mathbb{P}(\sum_{2 \leq j \leq n} x_j < an) \mathbb{P}(\sum_{j=1}^n t_j < an). \end{aligned}$$

Für hinreichend großes  $n$  sind die beiden letzten Faktoren  $\geq 1/2$  (nach dem Gesetz der großen Zahlen), also ist für die Endlichkeit von  $\mathbb{E}(\tilde{w})$

notwendig, dass

$$\sum_n n\mathbb{P}(x > nc)$$

endlich ist. Dies ist genau dann der Fall, wenn  $\mathbb{E}(x^2)$  endlich ist. Diese Bedingung ist auch hinreichend. Dazu nehmen wir zuerst an, dass auch  $\mathbb{E}(t^2) < \infty$  gilt und beginnen mit der Rekursion für die Wartezeit:

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

Quadrieren liefert

$$w_{n+1}^2 = (w_n + u_n)_+^2 \leq (w_n + u_n)^2.$$

Die Folge  $\mathbb{E}(w_n^2) = \mathbb{E}(\tilde{w}_n^2)$  ist nichtfallend und reell, also haben wir

$$\mathbb{E}(w_n^2) \leq \mathbb{E}(w_{n+1}^2) \leq \mathbb{E}(w_n^2) + 2\mathbb{E}(w_n)\mathbb{E}(u_n) + \mathbb{E}(u_n^2),$$

und schließlich

$$\mathbb{E}(w_n) \leq \frac{\mathbb{E}(u^2)}{-2\mathbb{E}(u)}.$$

Damit gilt dann auch

$$\mathbb{E}(\tilde{w}) = \lim_{n \rightarrow \infty} \mathbb{E}(w_n) \leq \frac{\mathbb{E}(u^2)}{-2\mathbb{E}(u)}.$$

Wenn  $\mathbb{E}(t^2)$  unendlich ist, dann können wir  $M > 0$  wählen und setzen  $t_n^* = \min(t_n, M)$ . Wenn  $M$  groß genug gewählt wird, gilt immer noch  $\mathbb{E}(t_n^*) > \mathbb{E}(x_n)$ . Dann gilt

$$u_n^* = x_n - t_{n+1}^* \geq u_n,$$

$$w_n^* \geq w_n$$

und

$$\tilde{w}^* \geq \tilde{w}.$$

$t_n^*$  ist aber beschränkt, und deswegen hat  $\tilde{w}^*$  und damit auch  $\tilde{w}$  endlichen Erwartungswert.

Generell ist das  $m$ -te Moment von  $\tilde{w}$  genau dann endlich, wenn  $x_n$  ein endliches  $(m+1)$ -tes Moment besitzt.

## 2.2 Der Satz von Little

Wir nehmen jetzt an, dass in unserer Schlange stationäres Verhalten herrscht und dass die wichtigen Größen (Zwischenankunftszeit, Bedienzeit, Wartezeit) endliche Erwartungswerte haben; diese Annahme benötigen wir eigentlich nur, damit Zeitmittel konvergieren. Wir wollen eine Beziehung zwischen der Ankunftsrate, der mittleren Anzahl der Kunden im System und der mittleren Aufenthaltsdauer finden.

Die Voraussetzungen sind

1.  $T_n/n = (t_1 + \dots + t_n)/n \rightarrow 1/\lambda$ ,
2.  $\bar{z}_n = (z_1 + \dots + z_n)/n \rightarrow \bar{z}$ .

Die Stationarität der Warteschlange hat (nach dem Birkhoff'schen Ergodensatz) zur Folge, dass  $\bar{z}$  existiert. Unter unseren üblichen Voraussetzungen ist der stationäre Prozess ergodisch und damit  $\bar{z}$  deterministisch. Das muss man nicht annehmen, die folgende Analyse funktioniert auch dann, wenn  $\bar{z}$  eine nichtdegenerierte Zufallsvariable ist. Wir nehmen aber in jedem Fall an, dass die Ankunftsrate  $\lambda$  deterministisch ist.

Wir nehmen an, dass wir jede Kundin für die Zeit, die sie im System verbringt, bezahlen müssen. Für die ersten  $n$  Kundinnen sind insgesamt  $z_1 + \dots + z_n$  Einheiten zu bezahlen. Andererseits sind pro Zeiteinheit  $N_t$  Einheiten zu bezahlen, also sind bis zum Zeitpunkt  $T$

$$\int_0^T N_t dt$$

Einheiten notwendig.  $z_1 + \dots + z_n$  liegt zwischen dem Betrag, der bis zur Ankunftszeit  $T_n$  aufgewendet wurde, und dem bis zum Abgang der ersten  $n$  Kunden  $T_n^* = \max_{i \leq n} (T_i + z_i)$ . Aus der Konvergenz von  $\bar{z}_n$  folgt  $z_n/n \rightarrow 0$  und auch  $\max_{i \leq n} z_i/n \rightarrow 0$ . Deshalb konvergiert sowohl  $T_n/n$  als auch  $T_n^*/n$  gegen  $1/\lambda$ . Für hinreichend großes  $n$  gilt

$$\frac{n(1 - \epsilon)}{\lambda} \leq T_n \leq T_n^* \leq \frac{n(1 + \epsilon)}{\lambda}.$$

Bis zum Zeitpunkt  $T_n$  ist höchstens  $z_1 + \dots + z_n$  ausgegeben worden, bis  $T_n^*$  mindestens so viel, also

$$\int_0^{n(1-\epsilon/\lambda)} N_t dt \leq z_1 + \dots + z_n \leq \int_0^{n(1+\epsilon/\lambda)} N_t dt.$$

Daraus folgt (für großes  $n$ )

$$\frac{\lambda}{n(1-\epsilon)} \int_0^{n(1-\epsilon)/\lambda} N_t dt \leq \frac{\lambda(z_1 + \dots + z_n)}{n(1-\epsilon)}$$

und

$$\frac{\lambda}{n(1+\epsilon)} \int_0^{n(1+\epsilon)/\lambda} N_t dt \leq \frac{\lambda(z_1 + \dots + z_n)}{n(1+\epsilon)}.$$

Daraus schließen wir

$$\frac{\lambda\bar{z}}{1+\epsilon} \leq \liminf_{T \rightarrow \infty} \frac{1}{T} \int_0^T N_t dt \leq \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T N_t dt \leq \frac{\lambda\bar{z}}{1-\epsilon},$$

und weil  $\epsilon > 0$  beliebig gewählt werden kann

$$\bar{N} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T N_t dt = \lambda\bar{z}.$$

Es existiert also das Zeitmittel  $\bar{N}$  der Anzahl der Kunden im System und ist gleich dem Produkt aus Ankunftsrate und mittlerer Aufenthaltsdauer.

Als spezielles Beispiel betrachten wir den Bedienplatz als eigenes System. Die Ankunftsrate ist dieselbe wie für das ganze System also  $\lambda$ , die Aufenthaltsdauer ist gleich der Bedienzeit. Als gilt für die Anzahl  $N$  der Kunden, die gerade bedient werden

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

Da  $N$  nur die Werte 0 und 1 annehmen kann, ist das gleich der Wahrscheinlichkeit, dass der Server besetzt ist, oder dem Anteil der Zeit (über ein langes Intervall betrachtet), den der Server beschäftigt ist (was ja mit der intuitiven Vorstellung von ‘‘Auslastung’’ zusammenpasst).

Ein weiteres Beispiel (nicht ganz ernst und leicht makaber): Wien hat traditionellerweise eine innige Beziehung zum Tod, es ist gewissermaßen eine große Warteschlange vorm Zentralfriedhof. Wir kennen die Schlangenlänge (sprich Einwohnerzahl — sagen wir 1.7 Millionen) und die mittlere Aufenthaltsdauer (die Lebenserwartung — sagen wir 80 Jahre). Wenn wir an die Stationarität dieser Schlange glauben, können wir damit die Ankunftsrate (die im stationären Fall mit der Abgangsrate übereinstimmt) als  $\lambda = \bar{N}/\bar{z} = 1700000/80 = 21250$  Personen pro Jahr bestimmen. Die exakten Zahlen für 2013 sind 1740000 für die Bevölkerung, 81.2 für die Lebenserwartung; die Anzahl der Sterbefälle in diesem Jahr war 16332, die der Geburten 18691. Die Schätzung aus dem Satz von Little ergibt

21428, ist also für beide Raten zu groß. Das liegt natürlich daran, dass dieses System eklatant nichtstationär ist. In diesem Licht ist es eigentlich erstaunlich, dass unsere Überlegungen immerhin einigermaßen die richtige Größenordnung geben.

# 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  $\mu$  weg (es kann ja niemand gehen). Somit gilt für

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

Wenn man  $p_n(t)$  auf die linke Seite bringt und durch  $\Delta t$  dividiert und  $\Delta t \rightarrow 0$  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

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

Die unbedingte Dichte ergibt sich also zu

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

$z$  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  $\beta$ ) ist. Wir erhalten also

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

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.

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

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

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

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

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

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  $\sigma$ . 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 vernünftig 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  $W(\cdot)$  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  $\Psi^+$  und  $\Psi^-$  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

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

oder

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

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

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

**Beispiel:**  $M/M/1$

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

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

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:

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

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  $n \rightarrow \infty$  ergibt sich

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

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

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

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

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

Wenn wir wieder die Erwartungswerte berechnen und  $n \rightarrow \infty$  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

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

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

$$\begin{aligned} \mathbb{E}(y^2) &\geq (\mathbb{E}(y))^2 = (\mathbb{E}(u))^2 \quad \text{also} \\ \mathbb{E}(w^2) &\leq \frac{\mathbf{Var}(u)}{-2\mathbb{E}(u)} = \frac{\mathbf{Var}(x) + \mathbf{Var}(t)}{-2\mathbb{E}(u)} . \end{aligned}$$

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

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

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  $n \rightarrow \infty$  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  $n \rightarrow \infty$  ist  $g(y) \rightarrow \mathbb{E}(u) < 0$ , es gibt also eine eindeutig bestimmte Lösung  $y_0$ , für die  $g(y_0) = y_0$ , und

$$\mathbb{E}(w) \geq g(y_0) .$$

## 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  $\Psi^+$  oder  $\Psi^-$  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

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

$[b > a]$ .

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  $s \rightarrow 0$  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

$$\Psi^+(s) \approx Cs(s - s_0)$$

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

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

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  $\lambda$  bzw.  $\mu$  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  $n$

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

$N_a(u)$  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  $\mu u$  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  $\Delta u$  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  $\Delta u$  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  $\Delta u$ :

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

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

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

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

$$F(x, 0) = 1 \quad (x \geq 0)$$

$$F(x, 0) = 0 \quad (x < 0)$$

$$F(0, u) = 0 \quad (u > 0) .$$

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:

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

also

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

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

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

Die Inversion der Laplace-Transformation liefert

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

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:

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

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  $\rho \rightarrow 1$  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

## Der Einfluss der Disziplin

### 5.1 Nichtpräemptive Disziplinen - FCFS und ihre Freundinnen

Die Fragestellung dieses Kapitels lässt sich am besten an einem Beispiel aus dem Alltag demonstrieren: in einem Supermarkt gibt es mehrere Kassen. Wir können entweder jeder Kasse eine eigene Schlange geben oder eine einzelne Warteschlange einrichten, von deren Kopf die Kundinnen zu den freiwerdenden Kassen geschickt werden. Welche dieser Möglichkeiten ist die bessere? An dieser Stelle wollen wir uns auf die Wartezeiten konzentrieren.

Eine einzelne Warteschlange implementiert offensichtlich die “first come first served” Disziplin für das ganze System, über alle Server hinweg. Unsere Frage dreht sich also darum, wie FCFS gegenüber anderen Disziplinen abschneidet.

Eine erste Feststellung ist, dass einzelne Schlangen für die einzelnen Server, in denen die Kundinnen einmal eingereiht werden und dann nicht mehr wechseln dürfen, zu einer größeren mittleren Wartezeit führen. Das liegt aber daran, dass einzelne Schlangen leer werden können während in anderen viele Kunden warten. Dieses Bild entspricht natürlich nicht der Praxis — in der Realität werden schon, wenn eine Schlange deutlich kürzer wird, Kundinnen hinüberwechseln. Wenn wir das vermeiden, also dafür sorgen, dass keine unnötigen Pausen entstehen, dann haben alle “einfachen” nichtpräemptiven Disziplinen dieselbe mittlere Wartezeit. Konkret nehmen wir von unseren Disziplinen und dem System an:

1. Die Bedienzeiten sind vertauschbar: für jede Permutation  $\pi$  hat



$(t_1, \dots, t_n, x_1, \dots, x_n)$  dieselbe Verteilung wie  $(t_1, \dots, t_n, x_{\pi(1)}, \dots, x_{\pi(n)})$ .

2. Es wird immer gearbeitet, wenn es Arbeit gibt.
3. Die Disziplin ist unabhängig von den Bedienzeiten — wir haben also keine Information über die Bedienzeiten, die wir für die Einteilung der Kundinnen verwenden können, und andererseits ändern sich die Bedienzeiten nicht, wenn sich die Wartezeit verändert.

Unter diesen Voraussetzungen ist die mittlere Wartezeit unabhängig von der Disziplin.

Zum Beweis stellen wir neben die Schlange mit einer beliebigen Disziplin eine FCFs-Schlange. In dieser sind die Ankunftszeiten dieselben wie in der originalen Schlange, die Bedienzeiten erhalten wir so, dass wir die Bedienzeit der  $n$ -ten Kundin der zweiten Schlange gleich der Bedienzeit des Kunden der ersten Schlange setzen, der als  $n$ -ter bedient wird. Dadurch haben beide Schlangen dieselben Ankunfts- und Abgangszeiten und daher zu jedem Zeitpunkt dieselbe Anzahl von Kundinnen  $N_t$ . Es sind also die mittleren Anzahlen von Kunden dieselben und damit nach dem Satz von Little auch die mittleren Wartezeiten.

Der Erwartungswert der Wartezeit gibt uns also keinen Anhaltspunkt zur Unterscheidung der einzelnen Disziplinen.

## 5.2 Präemptive Disziplinen — 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  $\lambda$ , 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  $u \rightarrow \infty$ . 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  $u \rightarrow 0$  ergibt sich

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

für  $u \rightarrow \infty$

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

## 5.3 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  $\lambda_i$  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 gegenüber 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.

### 5.3.1 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  $W_i$ ,  $i = 1, \dots, p$ , 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  $n \rightarrow \infty$  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.

### 5.3.2 Eine Methode zur Bestimmung der Wartezeit

Wir betrachten einen Kunden aus der Gruppe  $i$ , der das System betritt:  $N_{ij} \dots$  mittlere Anzahl der Kunden aus Gruppe  $j$ , die unser Kunde im System antrifft, und die vor ihm bedient werden (ausgenommen der Kunde, der eventuell gerade bedient werden, wenn unser Kunde ankommt).  $M_{ij} \dots$  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

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

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

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

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

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.

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

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 Transformaten.
4. Weiters sei  $N_t$  ein Poissonprozess (d.h. eine Folge von Ereignissen, wobei die Zeit zwischen zwei Ereignissen nach  $M_\lambda$  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  $\lambda t$ :

$$\mathbb{P}(N_T = n | T = t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} ,$$

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.

# Index

- Ankunftsrate, 9
- Ankunftszeit, 4
- Auslastung, 9
- Bedienrate, 9
- Bedienzeit, 4
- Disziplin, 4
- erzeugende Funktion, 40
- FCFS, 4
- Gedächtnislosigkeit, 6
- Kendall-Notation, 5
- Klassen, 36
- Kurznotation nach Kendall, 5
- Laplace - Transformierte, 40
- LCFS, 5
- Poissonprozess, 41
- Pollaczek - Khinchin Formel, 17
- präemptiv, 5
- Prioritäten, 36
- Rechenzeit, 32
- Round Robin, 34
- Satz
  - LIIOUVILLE, 21
  - Zentraler Grenzwertsatz für Warteschlangen, 27
- Spektralzerlegung, 20
- stationäre Verteilung, 8
- Verteilung
  - Allgemeine, 6
  - Deterministische, 6
  - Erlang-, 5
  - Exponential-, 5
  - hyperexponentielle, 5
  - stationäre, 8
- Wartezeit, 6
- Zeitscheibe, 32
- Zwischenankunftszeit, 4