Warteschlangentheorie und Callcenter

Vortrag im Rahmen der Lehrerfortbildung
„Stochastik und Matrizen: von Markov-Ketten bis zu Callcentern"
23. September 2009

Dr. Alexander Herzog,
Institut für Mathematik, TU Clausthal

1 Warteschlangenmodelle

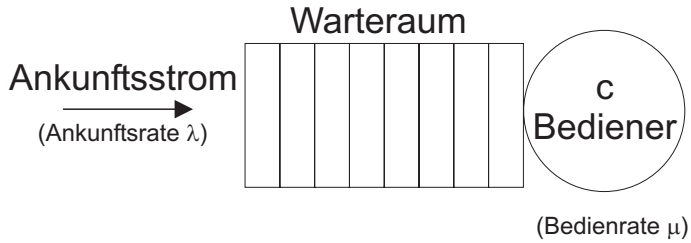

Abbildung 1: Allgemeines Warteschlangenmodell

Das Grundmodell in Abbildung 1 besteht aus folgenden Komponenten:

- **Ankunftsstrom:**
  Der Ankunftsstrom wird durch die zufälligen Ankunftszeitpunkte der Kunden am System charkterisiert und wird durch die Verteilung der Zeiten zwischen zwei Kundenankünften, den sogenannten **Zwischenankunftszeiten**, beschrieben. Es sei I die Zufallsvariable der Zwischenankunftszeiten, dann wird $\lambda := \frac{1}{\mathbf{E}[I]}$ die **Ankunftsrate** pro Zeiteinheit genannt. Es wird angenommen, dass die aufeinanderfolgenden Zwischenankunftszeiten stochastisch unabhängig und identisch verteilt sind.

- **Warteraum:**
  Vor dem Bedienschalter befindet sich ein Warteraum, in dem sich die eingetroffenen Kunden aufhalten, bevor sie bedient werden. Der Warteraum kann eine endliche oder eine unendliche Kapazität besitzen. Besitzt der Warteraum nur eine endliche Kapazität, so werden eintreffende Kunden abgewiesen, sobald alle Warteplätze belegt sind.

- **Bedienprozess:**
  An dem Bedienschalter arbeiten eine gewisse Anzahl an Bedienern parallel nebeneinander. Die **Bedienzeiten** der aufeinanderfolgenden Kunden bilden eine Folge stochastisch unabhängiger und identisch verteilter Zufallsvariablen mit Verteilungsfunktion F_S (S von englisch *Service*) und Erwartungswert $\mathbf{E}[S]$. $\mu := \frac{1}{\mathbf{E}[S]}$ entspricht der **Bedienrate** pro Zeiteinheit eines Bedieners.

- **Bedienregeln:**
  Über die sogenannten Bedienregeln wird festgelegt, nach welchem Prinzip die Kunden aus dem Warteraum zu den Bedienern geleitet werden. Da in den meisten Fällen der Ankunftszeitpunkt für die Priorität des Kunden im Warteraum verantwortlich ist, spricht man auch davon, dass sich vor dem Bedienschalter eine Warteschlange bildet.

1 Definition (Auslastung, Arbeitslast):
In einem System mit $c \in \mathbb{N}$ Bedienern definiert $\rho := \frac{\lambda}{c\mu}$ die Systemauslastung, wobei μ die Bedienrate eines einzelnen Bedieners pro Zeiteinheit darstellt. Außerdem wird $a := \frac{\lambda}{\mu}$ die **Arbeitslast** genannt.

Die Arbeitslast wird auch häufig zu Ehren von Agner Krarup Erlang, dem Begründer der Warteschlangentheorie, mit der Einheit `Erl` (für ,,Erlang") versehen, obwohl sie als Quotient zweier Raten einheitenlos ist.

2 Bezeichnung (Zustandswahrscheinlichkeit):
Im Folgenden sei π_n, $n \in \mathbb{N}$, die Wahrscheinlichkeit dafür, dass sich in einem Warteschlangensystem im stationären Zustand n Kunden im System, d.h. in der Warteschlange oder in Bedienung, befinden.

3 Definition (mittlere Anzahl an Kunden im System):
Die mittlere Anzahl an Kunden im System in einem Warteschlangensystem wird als

$$\mathbf{E}[N] := \sum_{n=0}^{\infty} n\pi_n$$

definiert. Besteht ein Warteschlangensystem aus c Bedienern, so definiert man die mittlere Anzahl an Kunden in der Warteschlange als

$$\mathbf{E}[N_Q] := \sum_{n=c+1}^{\infty} (n-c)\pi_n.$$

4 Bezeichnung (mittlere Wartezeit, mittlere Verweildauer):
In einem Warteschlangensystem wird mit W die Zufallsvariable der Wartezeit der Kunden beschrieben und mit V die Zufallsvariable der Verweilzeit der Kunden im System, d.h. der Summe aus Warte- und Bedienzeit.

5 Satz (John D. C. Little, 1961):
Für Warteschlangensysteme mit allgemeinen Zwischenankunfts- und Bedienzeiten sowie c parallelen Bedienern gilt im stationären Zustand folgender Zusammenhang zwischen der mittleren Warteschlangenlänge $\mathbf{E}[N_Q]$ und der mittleren Wartezeit der Kunden $\mathbf{E}[W]$:

$$\mathbf{E}[N_Q] = \lambda \mathbf{E}[W]$$

und folgender Zusammenhang zwischen der mittleren Anzahl an Kunden im System $\mathbf{E}[N]$ und der mittleren Verweilzeit der Kunden $\mathbf{E}[V]$:

$$\mathbf{E}[N] = \lambda \mathbf{E}[V].$$

2 Die Kendall-Notation

Viele Warteschlangenmodelle lassen sich nach der sogenannten Kendall-Notation (1953 von David George Kendall eingeführt) klassifizieren. Nach der Kendall-Notation wird einem Warteschlangensystem eine Buchstabenkombination der folgenden Form zugeordnet:

A / S / c / [C] [+ W] / [p] / [D]

Eckige Klammern deuten dabei an, dass der jeweilige Parameter optional ist. Die Platzhalter ,,A", ,,S", ,,c", ,,C", ,,W", ,,p" und ,,D" haben dabei diese Bedeutungen:

- **,,A"** gibt die Verteilung der Zwischenankunftszeiten (arrival) an.
- **,,S"** gibt die Verteilung der Bedienzeiten (service) an.
- **,,c"** gibt die Anzahl der identischen Bediener an. Es ist $c \in \mathbb{N}$. Übliche Werte sind ,,1" oder ,,c" um anzudeuten, dass eine feste Anzahl $c \geq 1$ an Bedienern zur Verfügung steht.
- **,,C"** gibt die maximale Anzahl an Kunden im System an (wartende Kunden sowie in Bedienung befindliche Kunden). Fehlt diese Angabe, so wird $C = \infty$ angenommen. Übliche Werte sind ,,c", ,,N" (mit $N > c$) und ,,∞".
- **W** gibt die Wartezeittoleranz-Verteilung der Kunden an. Fehlt diese Angabe, so besitzen alle Kunden ein unendliche Wartezeittoleranz und brechen einen Wartevorgang nie ab.

- **„p"** gibt die Populationsgröße, d.h. die Gesamtzahl an Kunden, die am System ankommen können, an. Fehlt diese Angabe, so wird $p = \infty$ angenommen.
- **„D"** beschreibt die Bedienregel (*service discipline*). Üblich sind hier insbesondere folgende Werte:
  - **FIFO** (*First in first out*) oder auch **FCFS** (*First come first served*): Bedienung der Kunden in der Ankunftsreihenfolge.
  - **LIFO** (*Last in first out*) oder auch **LCFS** (*Last come first served*): Bedienung der Kunden in umgekehrter Ankunftsreihenfolge, d.h. es wird immer der zuletzt eingetroffene Kunde als nächstes bedient.
  - **SPT** (*Shortest processing time*): Es wird immer derjenige wartende Kunde als nächstes bedient, dessen Bedienzeit am kürzesten ist.
  - **RANDOM** (zufällig) oder auch **SIRO** (*Serve in random order*): Der nächste zu bedienende Kunde wird jeweils zufällig ausgewählt.

  Wird keine Bedienregel angegeben, so wird stets FIFO angenommen.

Für die unterschiedlichen Verteilungstypen existieren folgende allgemein übliche Kürzel:

Symbol	Verteilung
M	Exponentialverteilung (Markov-Eigenschaft)
D	konstante Zwischenankunfts- bzw. Bedienzeiten (deterministisch)
E_k	Erlang-k-Verteilung
G	unbekannte Verteilung (generell), es sind nur Kenngrößen bekannt
H_k	Hyperexponentialverteilung (Linearkombination von k Exponentialverteilungen)
PH	Phasentypverteilung (Verteilung der Zeit, in der ein Markov-Prozess einen absorbierenden Zustand erreicht)

3 Unabhängigkeit der Kundenankünfte

Wenn sich die Kunden in Bezug auf Ihre Ladenbesuche nicht absprechen und auch nicht z.B. busladungsweise eintreffen, kann davon ausgegangen werden, dass die Kunden unabhängig von einander eintreffen. Betrachtet man ein solches System zum einem Zeitpunkt $t_0 \in \mathbb{R}_0^+$, so ist die Wahrscheinlichkeit, dass innerhalb der nächsten 5 Minuten ein Kunde eintrifft, unabhängig davon, ob innerhalb der letzten 5 Minuten ein Kunde eingetroffen ist oder nicht. Als bedingte Wahrscheinlichkeit formuliert bedeutet dies:

$$P(X \leq t_1 | X > t_0) = P(X \leq t_1 - t_0), \tag{1}$$

wobei $t_1 > t_0$ sei. Im einzelnen ist $P(X \leq t_1 | X > t_0)$ die Wahrscheinlichkeit, dass die Ankunft eines Kunden vor oder zu dem Zeitpunkt t_1 stattfindet ($X \leq t_1$), unter dem Vorauswissen, dass bis zum Zeitpunkt t_0 noch kein Kunde eingetroffen ist ($X > t_0$) und $P(X \leq t_1 - t_0)$ ist die Wahrscheinlichkeit, dass ein Kunde in der Zeitdifferenz zwischen t_0 und t_1 eingetroffen ist (ohne ein Vorauswissen über die Zeit zwischen 0 und t_0.

6 Definition (Bedingte Wahrscheinlichkeiten):
Es sei $P(B) > 0$, d.h. das Ereignis B trete mit einer Wahrscheinlichkeit größer als 0 ein. Dann sei $P(A|B)$ die bedingte Wahrscheinlichkeit dafür, dass A eintritt, wenn bekannt ist, dass B eingetreten ist. Es gilt

$$P(A|B) := \frac{P(A \cap B)}{P(B)},$$

wobei $P(A \cap B)$ die Wahrscheinlichkeit dafür ist, dass A und B gleichzeitig eintreten.

7 Definition (Exponentialverteilung):
Die Wahrscheinlichkeitsverteilung mit der Dichte

$$f_\lambda(x) := \begin{cases} 0 & \text{für } x < 0 \\ \lambda e^{-\lambda x} & \text{für } x \geq 0 \end{cases}$$

und der Verteilungsfunktion

$$F_\lambda(x) := \begin{cases} 0 & \text{für } x < 0 \\ 1 - e^{-\lambda x} & \text{für } x \geq 0 \end{cases}$$

heißt Exponentialverteilung mit Parameter $\lambda > 0$.

8 Satz:
Die Exponentialverteilung erfüllt (1).

Beweis:

$$\begin{aligned} P(X \leq t_1 | X > t_0) &= \frac{P(\{X \leq t_1\} \cap \{X > t_0\})}{P(X > t_0)} = \frac{P(t_0 < X \leq t_1)}{P(X > t_0)} = \frac{F(t_1) - F(t_0)}{1 - F(t_0)} \\ &= \frac{1 - e^{-\lambda t_1} - (1 - e^{-\lambda t_0})}{1 - (1 - e^{-\lambda t_0})} = \frac{e^{-\lambda t_0} - e^{-\lambda t_1}}{e^{-\lambda t_0}} = \frac{1 - e^{-\lambda t_1}/e^{-\lambda t_0}}{1} \\ &= 1 - e^{-\lambda(t_1 - t_0)} = F(t_1 - t_0) = P(X \leq t_1 - t_0). \end{aligned}$$

■

9 Bemerkung:
Es lässt sich auch die umgekehrte Richtung zeigen, also dass die Exponentialverteilung die einzige kontinuierliche Verteilung ist, die (1) erfüllt. Der Nachweis hierfür ist jedoch deutlich aufwendiger.

4 Modellierung eines Warteschlangensystems als Markovkette

Warteschlangensysteme mit exponentiell verteilten Zwischenankunfts- und Bedienzeiten lassen sich als Geburts- und Todesprozesse auffassen. Die Ankunft eines Kunden stellt eine Geburt dar und der Abschluss einer Bedienung einen Tod. Da Kunden auch in einem leeren System eintreffen können, handelt es sich um einen Geburts- und Todesprozess mit Einwanderung.

4.1 Zustandsabhängige Zwischenankunfts- und Bedienzeiten

Betrachtet werden soll ein M/M/c/K-Bediensystem, d.h. ein System mit exponentiell verteilten Zwischenankunfts- und Bedienzeiten, c parallelen Bedienern und K Warte- und Bedienplätzen. Die Ankunftsrate, d.h. die mittlere Anzahl an Ankünften pro Stunde, betrage λ, die mittlere Bedienrate pro Stunde μ. Für die zustandabhängigen Ankunfts- und Bedienraten ergeben sich damit:

$$\lambda_i := \begin{cases} \lambda & \text{für} \quad 0 \leq i < K \\ 0 & \text{für} \quad i = K \end{cases} \quad \text{und} \quad \mu_i := \begin{cases} i\mu & \text{für} \quad i < c \\ c\mu & \text{für} \quad i \geq c. \end{cases}$$

Befinden sich bereits K Kunden im System, so können keine weiteren Kunden mehr eintreffen, was zu $\lambda_K := 0$ führt. Sind weniger als c Kunden im System, so werden diese alle gleichzeitig bedient ($\lambda_i := i\mu$ für $i = 0...c-1$). Befinden sich jedoch c oder mehr Kunden im System, so sind alle Bediener ausgelastet und es werden genau c Kunden gleichzeitig bedient ($\lambda_i := c\mu$ für $i = c...K$).

4.2 Aufstellen der q-Matrix

Die q-Matrix beschreibt die Intensitäten, mit denen Übergänge von einem Zustand zu einem anderen stattfinden. Die Zeile gibt dabei jeweils den Ausgangszustand an und die Spalte den Zielzustand. Für den Fall des M/M/c/K-Systems ergibt sich:

$$Q := \begin{bmatrix} -\lambda_0 & \lambda_0 & & & & & & 0 \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & & & & \\ & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & & & & \\ & & \mu_3 & -(\lambda_3+\mu_3) & \lambda_3 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \mu_{K-1} & -(\lambda_{K-1}+\mu_{K-1}) & \lambda_{K-1} \\ 0 & & & & & \mu_K & -\mu_K \end{bmatrix}$$

5 Berechnung der stationären Verteilung und der stationären Wartezeitverteilung

5.1 Voraussetzungen für die Existenz einer stationären Verteilung

Eine stationäre Verteilung, die unabhängig vom Startzustand des Systems ist, kann sich nur dann einstellen, wenn jeder Zustand von jedem aus erreichbar ist (evtl. über mehrere Zwischenschritte). Der Zustandsraum wird in diesem Fall **irreduzibel** genannt.

Desweiteren muss die q-Matrix **konservativ** sein: Die Zeilensumme muss in allen Zeilen 0 betragen, d.h. die Intensität, mit der ein Zustand verlassen wird (negativer Wert in der Hauptdiagonale) muss genau den Intensitäten entsprechen, mit denen andere Zustände angenommen werden.

Schwieriger nachzuweisen ist die ebenfalls notwendige **Regularität**. Eine q-Matrix heißt regulär, wenn der zugehörige Feller-Prozess der zeitabhängigen Übergangswahrscheinlichkeiten stets eine 1 als Zeilensumme besitzt. Der Nachweis erfolgt allgemein über das Kriterium von Reuter. Für Geburts- und Todesprozesse gibt es davon abgeleitet spezielle Kriterien, so ist ein Geburts- und Todesprozess z.B. dann regulär, wenn der Zustandraum endlich ist oder stets $\lambda_i < \mu_i$ gilt. (Im vorliegenden Fall ist sogar beides erfüllt.)

Schließlich muss der Prozess noch **ergodisch** sein, damit eine stationäre Verteilung existieren kann. Ergodisch bedeutet, dass der Prozess mit einer positiven Wahrscheinlichkeit nach einer endlichen Zeit in seinen Startzustand zurückkehrt. Anschaulich bedeutet dies, dass das System nicht explodieren darf. Durch die Voraussetzung $\lambda_i < \mu_i$ ist dies bei uns gewährleistet.

5.2 Bestimmung der stationären Verteilung

Betrachtet man die Übergänge in einem kleinen Zeitschritt Δt, so sind praktisch nur direkte Übergänge nach links oder rechts in der Markovkette möglich, d.h. ist Δt sehr kurz, so kann nur ein Kunde ankommen oder ein Kunde das System verlassen (oder natürlich es passiert überhaupt nichts). Die Wahrscheinlichkeiten für die Ankunft bzw. den Abgang mehrerer Kunden wird mit dem Landau-Symbol $o(\Delta t)$ bezeichnet und fällt beim Grenzübergang $\Delta t \to 0$ weg.

Für den Zustand n und den Zeitschritt vom Zeitpunkt t zum Zeitpunkt $t + \Delta t$ ergibt sich:

$$\begin{aligned} \pi_n(t+\Delta t) &= \pi_n(t)[1-\lambda_n\Delta t-\mu_n\Delta t+o(\Delta t)] \\ &\quad +\pi_{n-1}(t)[\lambda_{n-1}\Delta t+o(\Delta t)] \\ &\quad +\pi_{n+1}(t)[\mu_{n+1}\Delta t+o(\Delta t)]. \end{aligned} \tag{2}$$

Die Wahrscheinlichkeit, sich zum Zeitpunkt $t + \Delta t$ im Zustand n ($n = 1...K-1$) aufzuhalten ($\pi_n(t + \Delta t)$), berechnet sich also aus der Wahrscheinlichkeit, zum Zeitpunkt t im Zustand n zu sein ($\pi_n(t)$) multipliziert mit der Wahrscheinlichkeit, diesen Zustand in der Zeit Δt nicht zu verlassen zuzüglich der Wahrscheinlichkeiten zum Zeitpunkt t in den Zuständen $n-1$ bzw. $n+1$ gewesen zu sein ($\pi_{n-1}(t)$ bzw. $\pi_{n+1}(t)$) multipliziert mit der Wahrscheinlichkeit, dass in der Zeit Δt ein Kunde eingetroffen ist bzw. fertig bedient wurde.

Zieht man $\pi_n(t)$ auf beiden Seiten von (2) ab und teilt durch Δt, so ergibt sich:

$$\frac{\pi_n(t+\Delta t)-\pi_n(t)}{\Delta t} = \pi_n(t)\frac{-\lambda_n\Delta t-\mu_n\Delta t+o(\Delta t)}{\Delta t}+\pi_{n-1}(t)\frac{\lambda_{n-1}\Delta t+o(\Delta t)}{\Delta t}+\pi_{n+1}(t)\frac{\mu_{n+1}\Delta t+o(\Delta t)}{\Delta t}.$$

Führt man nun den Grenzübergang $\Delta t \to 0$ durch, so ergibt sich:

$$\pi_n'(t) = \pi_n(t)(-\lambda_n-\mu_n)+\pi_{n-1}(t)\lambda_{n-1}+\pi_{n+1}(t)\mu_{n+1}.$$

Analog ergibt sich für die Randpunkte $n = 0$ und $n = K$:

$$\begin{aligned} \pi_0'(t) &= -\lambda_0\pi_0(t)+\mu_1\pi_1(t), \\ \pi_K'(t) &= \lambda_{K-1}\pi_{K-1}(t)-\mu_K\pi_K(t). \end{aligned}$$

Im eingeschwungenen Zustand ändern sich die Wahrscheinlichkeiten, mit denen man sich in einem bestimmten Zustand befindet, nicht mehr, d.h. es gilt $\pi_n'(t) = 0$ für $n = 0...K$ und $t \to \infty$. Damit ist also folgendes Gleichungssystem zu lösen:

$$\begin{aligned} -\lambda_0\pi_0+\mu_1\pi_1 &= 0, \\ \lambda_{n-1}\pi_{n-1}-(\lambda_n+\mu_u)\pi_n+\mu_{n+1}\pi_{n+1} &= 0, \quad n=1,\ldots,K-1, \\ \lambda_{K-1}\pi_{K-1}-\mu_K\pi_K &= 0. \end{aligned} \tag{3}$$

Aus der ersten Zeile folgt sofort

$$\pi_1 = \frac{\lambda_0}{\mu_1}\pi_0 \tag{4}$$

und aus der zweiten Zeile folgt für $n = 1$

$$\pi_2 = \pi_1\frac{\lambda_1+\mu_1}{\mu_2}-\pi_0\frac{\lambda_0}{\mu_2}. \tag{5}$$

Löst man (4) nach π_0 auf und setzt dies in (5) ein, so ergibt sich:

$$\pi_2 = \pi_1\frac{\lambda_1+\mu_1}{\mu_2}-\pi_1\frac{\mu_1}{\lambda_0}\frac{\lambda_0}{\mu_2} = \frac{\lambda_1}{\mu_2}\pi_1,$$

was die Induktionsannahme

$$\pi_n = \frac{\lambda_{n-1}}{\mu_n}\pi_{n-1} \tag{6}$$

nahelegt. Löst man die zweite Zeile von (3) nach $n+1$ auf und nimmt an, dass die Induktionsannahme für $1...n$ gilt, so ergibt sich:

$$\pi_{n+1} = \pi_n\frac{\lambda_n+\mu_n}{\mu_{n+1}}-\pi_{n-1}\frac{\lambda_{n-1}}{\mu_{n+1}} = \pi_n\frac{\lambda_n+\mu_n}{\mu_{n+1}}-\pi_n\frac{\mu_n}{\lambda_{n-1}}\frac{\lambda_{n-1}}{\mu_{n+1}} = \pi_n\frac{\lambda_n}{\mu_{n+1}}.$$

Durch rekursive Anwendung von (6) ergibt sich auch sofort:

$$\pi_n = \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0,$$

was auch für die letzte Zeile von (3), die bisher noch nicht berücksichtigt wurde, gilt.

Damit lassen sich jetzt alle Zustandswahrscheinlichkeiten π_n, $n = 1...K$, in Abhängigkeit von π_0 ausdrücken. Da jedoch $\sum_{n=0}^{K} \pi_n = 1$ gelten muss, lässt sich auch π_0 explizit bestimmen:

$$1 = \sum_{n=0}^{K} \pi_n = \pi_0 + \sum_{n=1}^{K} \pi_n = \pi_0 + \sum_{n=1}^{K} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0 = \pi_0 \left(1 + \sum_{n=1}^{K} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right).$$

Löst man diese Gleichung nach π_0 auf, so ergibt sich:

$$\pi_0 = \left(1 + \sum_{n=1}^{K} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1}.$$

Insgesamt ergeben sich damit für die stationären Zustandswahrscheinlichkeiten:

$$\pi_n = \begin{cases} \left(1 + \sum_{m=1}^{K} \prod_{i=1}^{m} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} & \text{für } n = 0, \\ \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0 & \text{für } n = 1...K. \end{cases}$$

5.3 Bestimmung der Verteilung der Anzahl an Kunden im System im stationären Fall

Wählt man $K = \infty$, d.h. lässt man die Warteraumbegrenzung weg, so lässt sich auch die Verteilung für die Anzahl an Kunden im System relativ einfach berechnen. (Ohne die Annahme $K = \infty$ würde die Γ-Funktion in dem Ergebnis auftreten.)

Die Wahrscheinlichkeitsverteilung $P(W \leq t)$ für die Wartezeiten der Kunden im stationären Fall lässt sich zunächst mit Hilfe des Satzes von der totalen Wahrscheinlichkeit wie folgt beschreiben:

$$P(W \leq t) = \sum_{n=0}^{\infty} P(W \leq t | N = n) \cdot \pi_n$$

Ist die Anzahl der Kunden im System kleiner oder gleich $c - 1$, so ist mindestens ein Bediener frei und ein eintreffender Kunde kann sofort bedient werden, d.h. für den Fall $n \leq c - 1$ gilt $P(W \leq t | N = n) = 0$ für jedes $t \geq 0$. Außerdem ist $\sum_{n=0}^{c-1} \pi_n$ gerade die Wahrscheinlichkeit dafür, dass ein neu eintreffender Kunde nicht warten muss, also $P(W = 0)$. Damit folgt:

$$\begin{aligned} P(W \leq t) &= \sum_{n=0}^{\infty} P(W \leq t | N = n) \cdot \pi_n = \sum_{n=0}^{c-1} 1 \cdot \pi_n + \sum_{n=c}^{\infty} P(W \leq t | N = n) \cdot \pi_n \\ &= P(W = 0) + \sum_{n=c}^{\infty} P(n - c + 1 \text{ Kunden bedient } \leq t | N = n) \cdot \pi_n. \end{aligned}$$

Im Fall $n \geq c$ sind alle Bediener ausgelastet, d.h. die Bedienrate pro Zeiteinheit lautet $c\mu$. Der Takt, mit dem die Warteschlange abgearbeitet wird, entspricht dem Minimum von c stochastisch unabhängigen und mit Parameter $\mu > 0$ exponentiell verteilten Zufallsgrößen. (Die Bedienzeiten der wartenden

Kunden sind exponentiell verteilt; die Restbedienzeiten der Kunden in Bedienung genügen aufgrund der Gedächtnislosigkeit der Exponentialverteilung ebenfalls der Exponentialverteilung.) Damit entspricht die Dauer für die Bedinung von $n-c+1$ Kunden einer Erlang$(n-c+1, c\mu)$ verteilten Zufallsgröße. Und es folgt:

$$P(W \leq t) = P(W=0) + \sum_{n=c}^{\infty} \int_0^t f_{\text{Erlang}(n-c+1,c\mu)}(x)dx \cdot \pi_n.$$

Mit $f_{\text{Erlang}(k,\lambda)}(x) := \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}$ für $x \geq 0$ folgt weiter:

$$P(W \leq t) = P(W=0) + \sum_{n=c}^{\infty} \int_0^t \frac{(c\mu)^{n-c+1}}{(n-c)!} x^{n-c} e^{-c\mu x} \pi_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} dx.$$

Für $n \geq c$ war

$$\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = \prod_{i=1}^{c} \frac{\lambda}{i\mu} \cdot \prod_{i=c+1}^{n} \frac{\lambda}{c\mu} = \frac{\lambda^c}{c!\mu^c} \cdot \left(\frac{\lambda}{c\mu}\right)^{n-c}.$$

Damit:

$$\begin{aligned}
P(W \leq t) &= P(W=0) + \pi_0 \frac{\lambda^c}{\mu^c c!} \sum_{n=c}^{\infty} \int_0^t \frac{(c\mu)^{n-c+1}}{(n-c)!} x^{n-c} e^{-c\mu x} \left(\frac{\lambda}{c\mu}\right)^{n-c} dx \\
&= P(W=0) + \pi_0 \frac{\lambda^c}{\mu^c c!} \int_0^t c\mu \cdot e^{-c\mu x} \underbrace{\sum_{n=c}^{\infty} \frac{\left(c\mu x \frac{\lambda}{\mu c}\right)^{n-c}}{(n-c)!}}_{=\sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} = e^{\lambda x}} dx \\
&= P(W=0) + \pi_0 \frac{\lambda^c}{\mu^c c!} \int_0^t c\mu \cdot e^{-c\mu x} \cdot e^{\lambda x} dx \\
&= P(W=0) + \pi_0 \frac{\lambda^c}{\mu^{c-1}(c-1)!} \left[-\frac{1}{c\mu - \lambda} e^{-(c\mu-\lambda)x}\right]_{x=0}^{x=t} \\
&= P(W=0) + \pi_0 \frac{\lambda^c}{\mu^{c-1}(c-1)!(c\mu-\lambda)} \left(1 - e^{-(c\mu-\lambda)t}\right).
\end{aligned}$$

Nun gilt aber:

$$\begin{aligned}
P(W>0) &= P(N \geq c) = \sum_{n=c}^{\infty} \pi_n = \sum_{n=c}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0 = \sum_{n=c} \underbrace{\prod_{i=1}^{c} \frac{\lambda}{i\mu}}_{=\frac{\lambda^c}{c!\mu^c}} \cdot \underbrace{\prod_{i=c+1}^{n} \frac{\lambda}{c\mu}}_{=\left(\frac{\lambda}{c\mu}\right)^{n-c}} \pi_0 \\
&= \frac{\lambda^c}{c!\mu^c} \pi_0 \sum_{n=1}^{\infty} \left(\frac{\lambda}{c\mu}\right)^n = \frac{\lambda^c}{c!\mu^c} \cdot \frac{1}{1-\frac{\lambda}{c\mu}} \pi_0 = \pi_0 \frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)}.
\end{aligned}$$

Und damit schließlich:

$$\begin{aligned}
P(W \leq t) &= P(W=0) + P(W>0)\left(1 - e^{-(c\mu-\lambda)t}\right) = 1 - P(W>0) + P(W>0)\left(1 - e^{-(c\mu-\lambda)t}\right) \\
&= 1 - P(W>0) \cdot e^{-(c\mu-\lambda)t}.
\end{aligned}$$

6 Berechnung weiterer Kenngrößen

Mit Hilfe der stationären Verteilung lassen sich eine Reihe von Kenngrößen berechnen. Dafür wird insbesondere die Formel von Little (siehe Satz 5) verwendet.

- Mittlere Anzahl an Kunden im System:

$$\mathrm{E}[N] = \sum_{n=0}^{K} n \cdot \pi_n = \sum_{n=0}^{K} n \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0.$$

- Mittlere Verweilzeit (Summe aus Wartezeit und Bedienzeit):

$$\mathrm{E}[V] = \frac{1}{\lambda}\mathrm{E}[N] = \frac{1}{\lambda} \sum_{n=0}^{K} n \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0.$$

- Mittlere Anzahl an Kunden in der Warteschlange:

$$\mathrm{E}[N_Q] = \sum_{n=c+1}^{K} (n-c) \cdot \pi_n = \sum_{n=c+1}^{K} (n-c) \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0.$$

- Mittlere Wartezeit der Kunden:

$$\mathrm{E}[W] = \frac{1}{\lambda}\mathrm{E}[N_Q] = \frac{1}{\lambda} \sum_{n=c+1}^{K} (n-c) \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \pi_0.$$

- Besetztwahrscheinlichkeit, d.h. die Wahrscheinlichkeit, dass ein neu eintreffender Kunde abgewiesen wird:

$$P(B) = \pi_K = \prod_{i=1}^{K} \frac{\lambda_{i-1}}{\mu_i} \pi_0.$$

- Wahrscheinlichkeit, dass ein Kunde warten muss (Blockierwahrscheinlichkeit):

$$P(W > 0) = \pi_0 \frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu - \lambda)}.$$

- Service-Level zum t-Sekunden-Niveau (Wahrscheinlichkeit, dass ein Kunde höchstens t-Sekunden warten muss):

$$P(W \leq t) = 1 - P(W > 0) \cdot e^{-(c\mu - \lambda)t}.$$

  Häufig soll ein System so ausgelegt werden, dass 80% der Kunden innerhalb von weniger als 20 Sekunden bedient werden oder sogar 90% in weniger als 10 Sekunden.