TU Clausthal

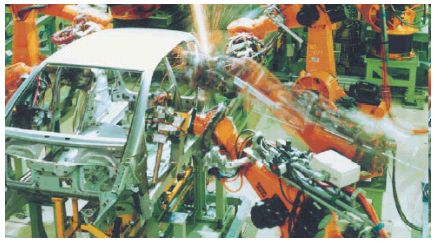
Warteschlangentheorie und Callcenter

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

A. Herzog,
Institut für Mathematik

TU Clausthal

Wo treten Warteschlangen auf und warum ?

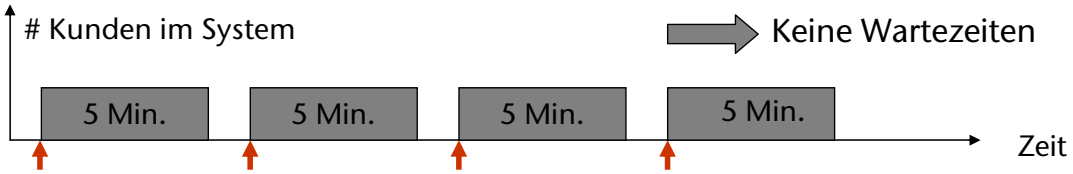


Kunden, Aufgaben, Werkstücke usw. treffen ein

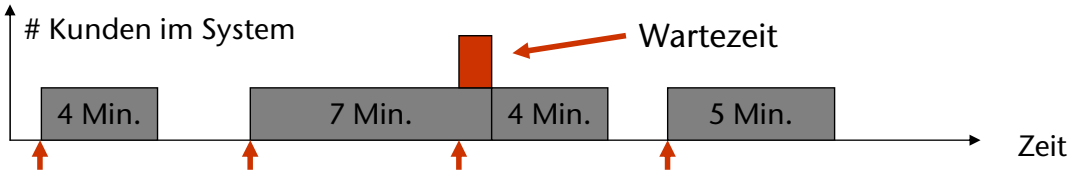
Bedienung bzw. Bearbeitung der Aufgabe

Ankünfte und Bedienungen

- Gleichmäßige Ankünfte, identische Bedienzeiten



- Gleichmäßige Ankünfte, **unterschiedliche** Bedienzeiten



- **Unterschiedliche** Zwischenankunftszeiten, unterschiedliche Bedienzeiten

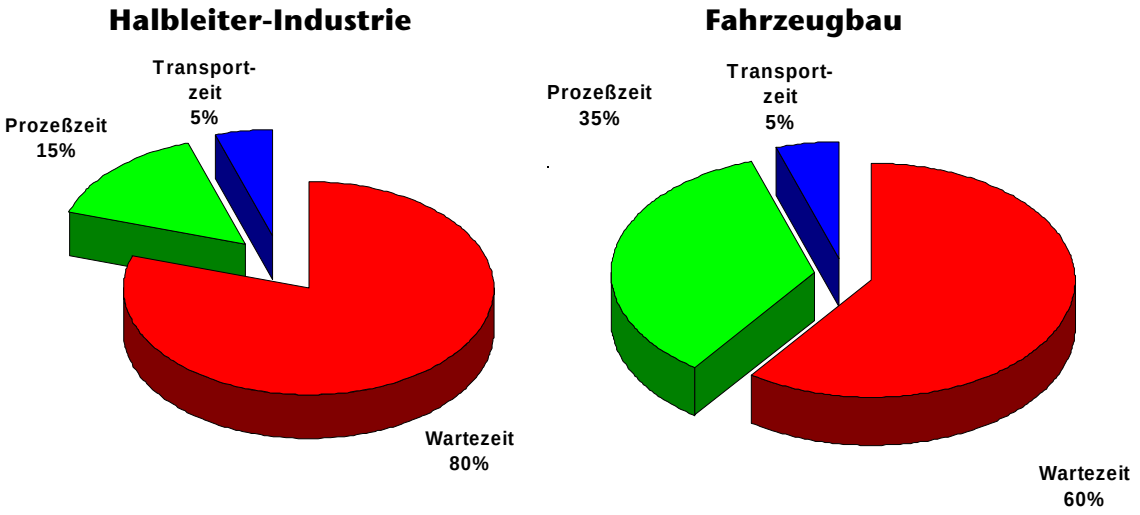
Wo treten in der Praxis Warteschlangen auf ?

- **Fertigungsstraßen**
  (Hohe Auslastung der Maschinen ⇔ Stauraum für halbfertige Bauteile)

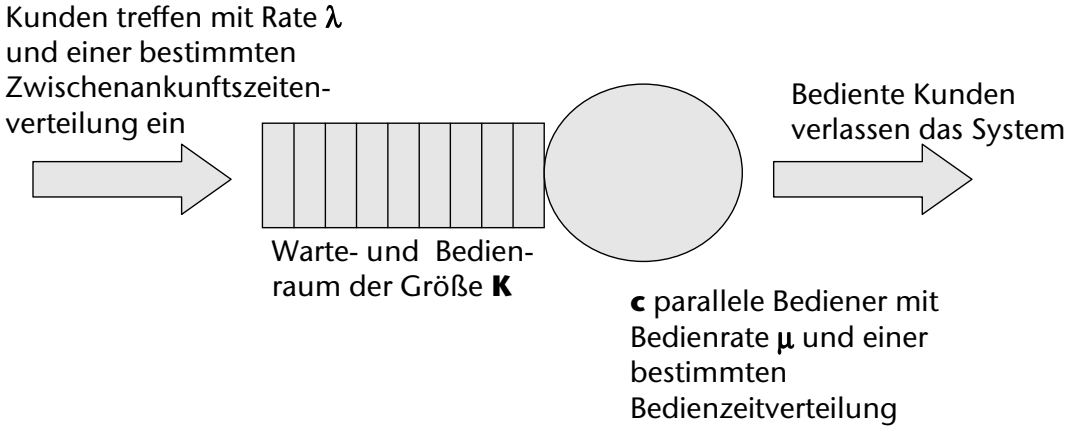
- **Kundenbediensysteme** (wie z.B. Hotlines)
  (Viele Mitarbeiter ⇔ Lange Wartezeiten der Kunden)

- **Flugverkehr**
  (Vollauslastung der Ressource „Rollbahn“⇔ Warteschleifen)

Haben die Wartezeiten in einem Produktionsprozess wirklich Einfluss auf die gesamte Produktionszeit ?



Mathematisches Modell eines Bediensystems



- Warteschlangen-Systeme werden nach der sogenannten **Kendall-Notation** beschrieben, z.B. **M/M/c/K**. Der erste Wert gibt den Typ der Zwischenankunftsverteilung an (hier M=exponentiell), der zweite Wert den Typ der Bedienzeitverteilungen und dritter und vierter Wert geben die Anzahl der Bediener bzw. die Größe von Warte- und Bedienraum an.

Erste elementare Erkenntnisse über Bediensysteme

- Ist **K** endlich, so bleibt die Anzahl an Kunden im System stets endlich.
- Ist **K** unendlich und $\lambda \geq \mu$, so wächst die Warteschlangenlänge unaufhaltsam, d.h. für ein stabiles System muss stets $\lambda < \mu$ gelten.
- Die mittlere Warteschlangenlänge ergibt sich als Produkt aus der Ankunftsrate λ und der mittleren Wartezeit. (Formel von Little, 1961)
- Treffen die Kunden unabhängig von einander ein, so lässt sich die mittlere Wartezeit aus Ankunftsrate, Bedienrate und Varianz der Bedienzeiten berechnen. (Pollaczek-Chinschin-Formel, 1930)

TU Clausthal

Verteilung der Zwischenankunftszeiten

- Im Folgenden wird angenommen, dass die Kunden stets **unabhängig** von einander eintreffen, d.h. dass gilt:

  $$P(X \leq t_1 | X > t_0) = P(X \leq t_1 - t_0)$$

  (Die Wahrscheinlichkeit, dass bis zum Zeitpunkt $\mathbf{t_1}$ ein Kunden eintrifft, wenn bis zum Zeitpunkt $\mathbf{t_0}$ noch kein Kunde angekommen ist, ist genau so hoch, wie die Wahrscheinlichkeit, dass in der Zeitspanne von $\mathbf{t_0}$ bis $\mathbf{t_1}$ ein Kunde ankommt – ohne das Vorwissen, dass in der Zeit von 0 bis $\mathbf{t_0}$ noch kein Kunde eingetroffen ist.)
- Die **Exponentialverteilung**

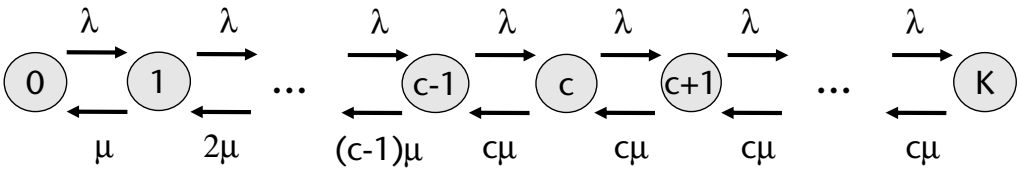
  $$f_\lambda(x) := -\lambda e^{-\lambda x}, \quad F_\lambda(x) := 1 - e^{-\lambda x}, \quad t \geq 0,\ \lambda > 0.$$

  erfüllt genau diese Voraussetzung.
- Diese Eigenschaft wird auch **Gedächtnislosigkeit** genannt.

Verteilung der Bedienzeiten

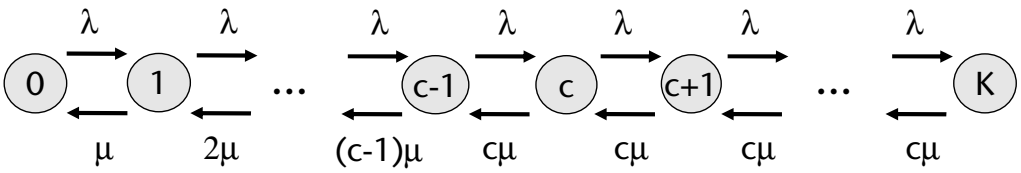
- Im Folgenden wird ebenfalls angenommen, dass die Bedienzeiten exponentiell verteilt seien.
- Bei den Zwischenankunftszeiten trifft die Annahme der Exponentialverteilung in der Regel zu; bei den Bedienzeiten ist dies nicht immer der Fall, d.h. hier entstehen Modellierungsfehler.
- Häufig nimmt man diese Abweichungen zwischen Modell und Realität jedoch in Kauf, da sich viele Kenngrößen nur unter der Annahme der Exponentialverteilung explizit berechnen lassen.

Mathematische Modellierung von Bediensystemen (1)



- Darstellung als sogenannter **Markov-Prozess**.
- Die **Kreise** stellen Zustände dar.
- Die **Verbindungspfeile** stellen mögliche Übergänge zwischen den Zuständen dar. Die Werte an den Pfeilen geben die Raten an, mit denen die jeweiligen Übergänge auftreten.
- **Markov-Eigenschaft**: Die Raten, mit denen ein Übergang in einen anderen Zustand erfolgen, hängen nur vom aktuellen Zustand und nicht von der Vorgeschichte ab. (Dies bedeutet, der Prozess ist gedächtnislos. Hier geht die Exponentialverteilungsannahme ein.)

Mathematische Modellierung von Bediensystemen (2)



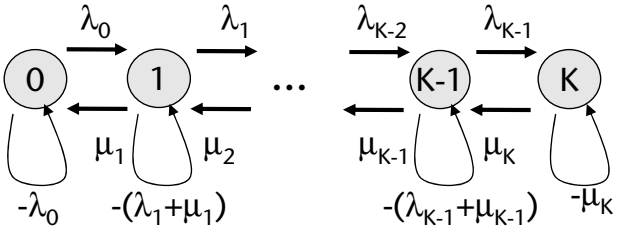
- Zur Vereinfachung der Schreibweise werden die Ankunfts- und Bedienraten mit λ_i und μ_i bezeichnet, wobei **i** die momentane Anzahl an Kunden im System darstellt.

- Damit gilt hier also:

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

TU Clausthal

Mathematische Modellierung von Bediensystemen (3)



- Die Übergangsraten von einem Zustand in einen anderen lassen sich als Matrix darstellen:

Eintrag in Zeile 2 und Spalte 3:
Rate, mit der Übergänge vom Zustand „1“ in den Zustand „2“ erfolgen.

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

Stationäre Verteilung

- Häufig interessiert man sich für die Wahrscheinlichkeitsverteilung der Anzahl an Kunden im System nach einer langen Laufzeit, d.h. wenn das System eingeschwungen ist.
- Aus dieser Verteilung lassen sich dann leicht Kenngrößen wie die **mittlere Anzahl an Kunden im System** oder die **mittlere Wartezeit** berechnen.
- Voraussetzung für die Existenz einer statonären Verteilung:
  Der Zustandsraum muss **irreduzibel** sein, d.h. jeder Zustand muss (ggf. über mehrere Schritte) von jedem aus erreichbar sein, und die q-Matrix muss **konservativ** (Zeilensumme=0), **regulär** (die Übergangswahrscheinlichkeiten müssen sich zu 1 aufsummieren) und **ergodisch** (hier gleichbedeutend mit $\lambda<\mu$) sein.

TU Clausthal

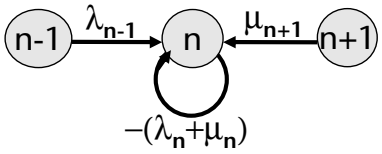
Bestimmung der stationären Verteilung (1)

- $\pi_n(t)$ sei die Wahrscheinlichkeit, dass sich der Prozess zum Zeitpunkt t im Zustand n befindet.
- Aufstellung eines Differenzengleichungssystem für $\pi_n(t+\Delta t)$, d.h. den Systemzustand nach einem kleinen Zeitschritt Δt.
- Grenzübergang $\Delta t \to 0$: Es entsteht ein Differentialgleichungssystem.
- Da sich die Wahrscheinlichkeiten im eingeschwungenen Zustand nicht mehr ändern, gilt $\pi' = 0$ und somit ergibt sich ein großes lineares Gleichungssystem.
- Bestimmung einer allgemeinen Lösung des unterbestimmten linearen Gleichungssystems.
- Da sich der Prozess stets in irgendeinem Zustand befinden muss und somit die Summe über alle Wahrscheinlichkeiten 1 beträgt, ergibt sich die endgültige Lösung des LGS durch Normierung.

Bestimmung der stationären Verteilung (2)

- Die zeitabhängigen Zustandswahrscheinlichkeiten $\pi_n(t)$:

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



- Im stationären Fall gilt $\pi' = 0$.

- Lineares Gleichungssystem zur Beschreibung des stationären Zustands:

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

Bestimmung der stationären Verteilung (3)

- Lösung des LGS ergibt zunächst:

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

d.h. das System ist einfach unterbestimmt.

- Es gilt jedoch $\sum_{n=0}^{K}\pi_n=1$. Damit lässt sich π_0 explizit berechnen:

$$1=\sum_{m=0}^{K}\pi_m=\pi_0+\sum_{m=1}^{K}\left[\prod_{i=1}^{m}\frac{\lambda_{i-1}}{\mu_i}\pi_0\right]=\pi_0\cdot\left[1+\sum_{m=1}^{K}\prod_{i=1}^{m}\frac{\lambda_{i-1}}{\mu_i}\right]$$

- Damit ergibt sich als endgültige Lösung:

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

Berechnung verschiedener Kenngrößen

- Mit Hilfe der stationären Verteilung π_n lassen sich direkt viele interessante Kenngrößen bestimmen:
  - Mittlere Anzahl an Kunden im System:
    $\mathrm{E}[N] = \sum_{n=0}^{K} n \cdot \pi_n$
  - Besetztwahrscheinlichkeit, d.h. die Wahrscheinlichkeit, dass neu eintreffende Kunden abgewiesen werden:
    $P(B) = \pi_K$
  - Mittlere Anzahl an Kunden in der Warteschlange:
    $\mathrm{E}[N_Q] = \sum_{n=c+1}^{K} (n-c) \cdot \pi_n$
  - Mittlere Wartezeit (folgt unmittelbar mit der Formel von Little):
    $\mathrm{E}[W] = \frac{1}{\lambda}\mathrm{E}[N_Q]$

Ursprünge der Warteschlangentheorie: Agner Krarup Erlang

- Dänischer Mathematiker, 1878-1929.
- Begründer der Warteschlangentheorie.
- Untersuchte Fragestellung zur Leistungsbemessung der manuellen Telefonvermittlungsstellen („Fräulein vom Amt“) für eine dänische Telefongesellschaft.

Agner Krarup Erlang

Bis heute relevante Ergebnisse:

- **Erlang-B-Formel** („Erlang-Verlustformel“): Bestimmung der Wahrscheinlichkeit, dass in einem System ohne Warteraum alle Leitungen belegt sind, d.h. ein neu eintreffender Anruf „verloren“ geht.
- **Erlang-C-Formel**: Wahrscheinlichkeitsverteilung für die Wartezeit eines Kunden in einem System mit Warteraum und **c** parallelen Bedienern.

Erlang-C Formel

- **Annahmen:** Der Warteraum ist unbeschränkt, die Zwischenankunfts-zeiten und die Bedienzeiten sind exponentiell verteilt, es gibt **c** parallel arbeitende Bediener und **a:**=λ/μ ist bekannt.

$$P_1 := \frac{\frac{a^c}{c!} \cdot \frac{c}{c-a}}{\left(\sum_{n=0}^{c-1} \frac{a^n}{n!}\right) + \frac{a^c}{c!} \cdot \frac{c}{c-a}}$$

$$P(W \leq t) = 1 - P_1 \cdot e^{-\mu(c-a)t}$$

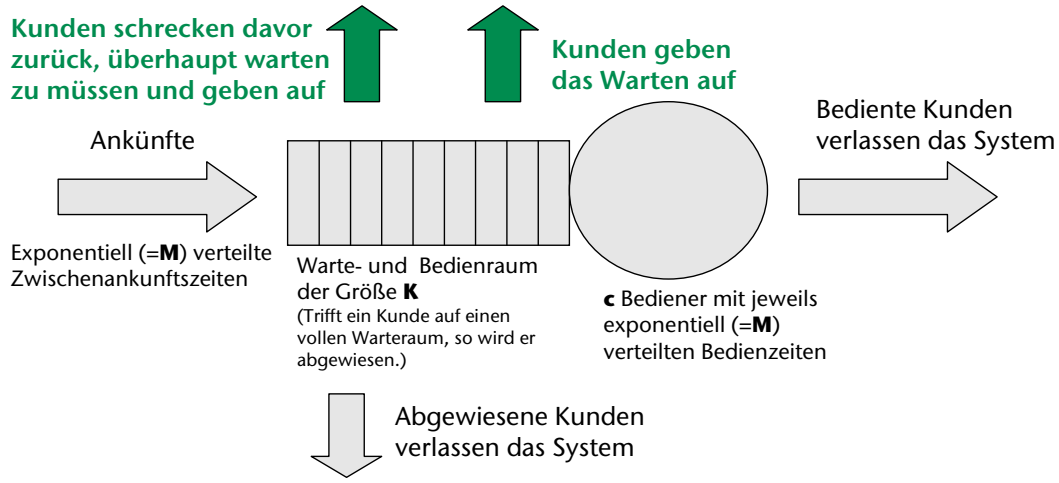
- **P(W≤t)** gibt die Wahrscheinlichkeit dafür an, dass die Wartezeit eines Kunden kleiner oder gleich **t** ist.
- Beweisidee:

$$P(W \leq t) = 1 \cdot P(N < c) + \sum_{n=c}^{\infty} \text{E[Bedienzeit für } n-(c+1)\text{] Kunden} \cdot \pi_n$$

Kritik am bisherigen Modell

- Das System bildet **c** unabhängige Bediener und einen begrenzten Warteraum (inkl. Abweisung von Kunden) ab.
- Die Annahme der Exponentialverteilung für die Zwischenankunftszeiten ist gerechtfertigt. Bei den Bedienzeiten kann der Fehler, der durch die Annahme der Exponentialverteilung auftritt, meist vernachlässigt werden.
- Das System kann ungeduldige Kunden nicht abbilden.
- Da es keine Abbrecher gibt, gibt es in diesem Modell auch keine Kunden, die später einen weiteren Versuch starten.

Ungeduldige Kunden



Die Wartezeittoleranz der Kunden sei exponentiell (=**M**) verteilt mit Abbruchrate **ν** und die Zurückscheurate betrage **β**.

→ **M/M/c/K+M - Modell**

Das M/M/c/K+M - Modell

- Die um zurückscheuende Kunden (→β) und Warteabbrecher (→ν) erweiterten Übergangsraten ergeben sich zu :

$$\lambda_i := \begin{cases} \lambda & \text{für} \quad 0 \le i < c \\ (1-\beta)\lambda & \text{für} \quad c \le i < K \\ 0 & \text{für} \quad i = K \end{cases} \quad \text{und} \quad \mu_i := \begin{cases} i\mu & \text{für} \quad i < c \\ c\mu + (i-c)\nu & \text{für} \quad i \ge c. \end{cases}$$

- Mit ähnlichen Überlegungen, wie im M/M/c/K – Fall ergibt sich dann:

$$P(W \le t) = 1 - C_K \pi_0 + \pi_0 \sum_{n=c}^{K-1} C_n \left(\frac{n! - \Gamma(n+1, (c\mu+\nu)t)}{(n-c!)(c\mu+\nu)^c} - 1 \right)$$

mit

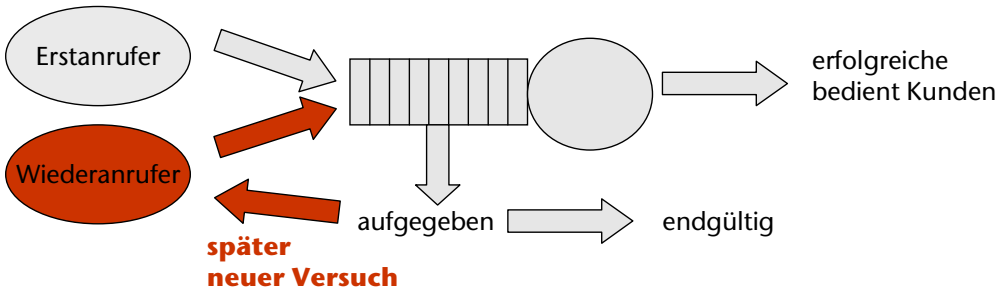
$$C_n := \frac{a^n (1-\beta)^{n-c}}{c! \prod_{i=1}^{n-c} \left(c + \frac{i\nu}{\mu}\right)}, \quad \pi_0 := \left(\sum_{n=0}^{c-1} \frac{a^n}{n!} + \sum_{n=c}^{K} C_n \right)^{-1} \text{ und}$$

$$\Gamma(x,z) := \int_z^\infty t^{x-1} e^{-t} dt, \;\; x \in \mathbb{R} \setminus \mathbb{Z}^-, \;\; z \in \mathbb{R}_0^+.$$

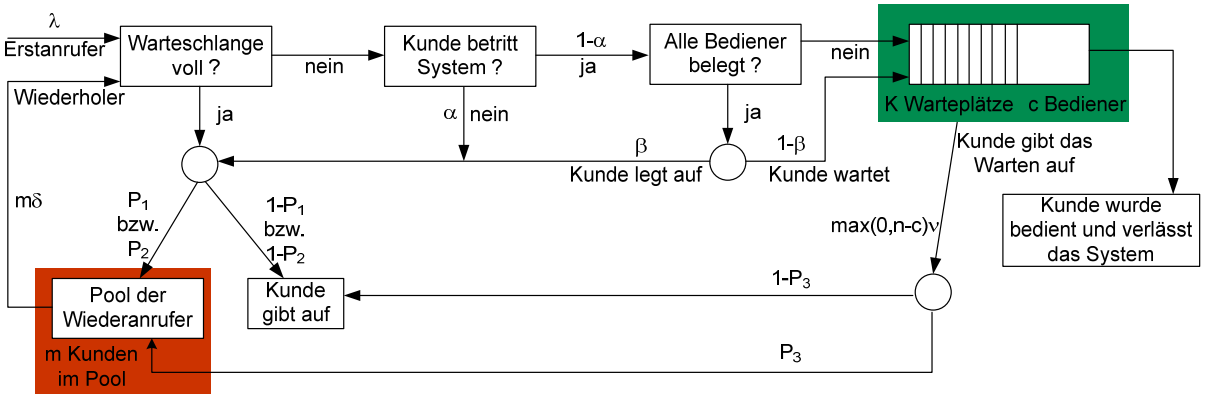
Kritik am M/M/c/K+M - Modell

- Durch die Modellierung von **Abbrechern** lässt sich der **Ist-Zustand** in einem Warteschlangensystem mit ungeduldigen Kunden gut nachbilden.

- Aussagen über die **Auswirkungen von Änderungen** am System lassen sich jedoch nicht treffen, da diese auch die Anzahl der **Wiederanrufer** und damit der Anrufer insgesamt ändern.



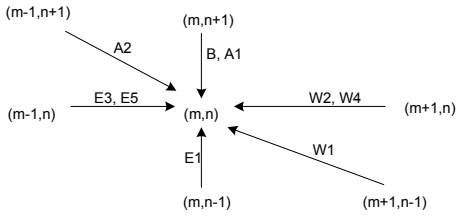
Modell mit Abbrechern und Wiederanrufern



Bisher: Eindimensionaler Zustandsraum *(Anzahl Kunden im System)*
Jetzt: Zweidimensionaler Zustandsraum
(Anzahl Kunden im Pool, Anzahl Kunden im System)

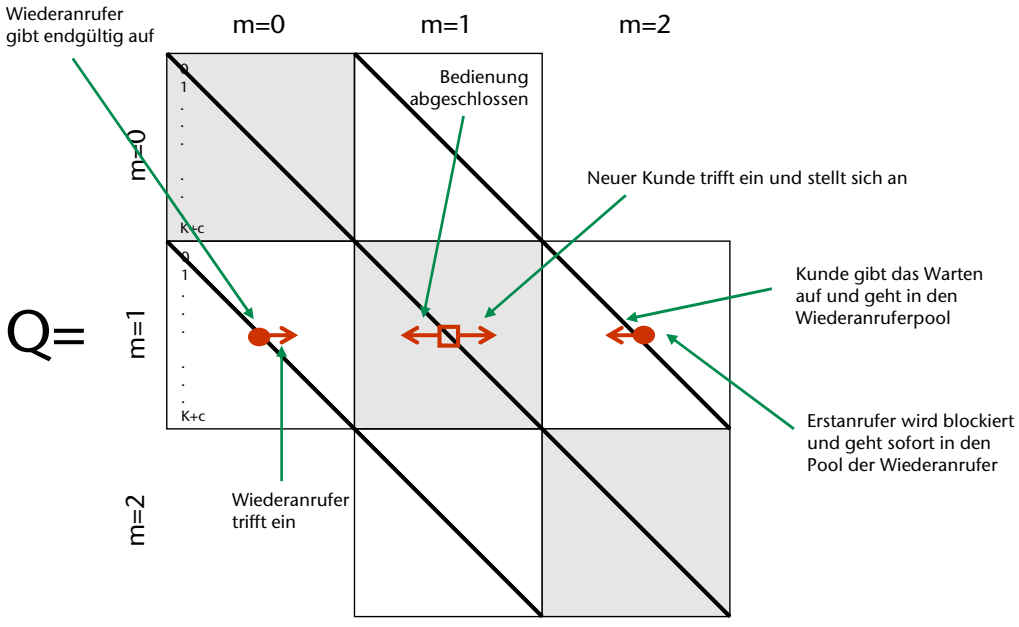
Mögliche Zustandsübergänge

(m=Kunden im Pool, n=Kunden im System)

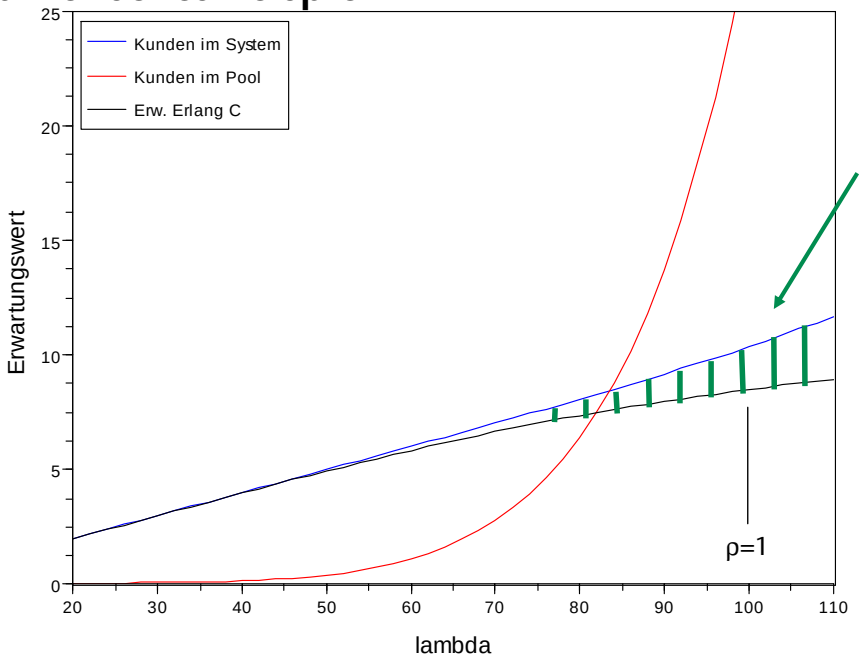


- **E1:** Ein Erstanrufer trifft ein und stellt sich an die Warteschlange an bzw. wird sofort bedient.
- **E2:** Ein Erstanrufer trifft ein, stellt fest dass es keinen freien Warteraum mehr gibt und verlässt das System endgültig. Oder aber: Ein Erstanrufer trifft ein, legt trotz vorhandenem Warteraum sofort wieder auf (verwählt usw.) und verlässt das System endgültig.
- **E3:** Ein Erstanrufer trifft ein, stellt fest dass es keinen freien Warteraum mehr gibt und begibt sich in den Pool der Wiederanrufer. Oder aber: Ein Erstanrufer trifft ein, legt trotz vorhandenem Warteraum sofort wieder auf (verwählt usw.) und begibt sich in den Pool der Wiederanrufer.
- **E4:** Ein Erstanrufer trifft ein, stellt fest dass er warten müsste und verlässt das System endgültig.
- **E5:** Ein Erstanrufer trifft ein, stellt fest dass er warten müsste und begibt sich in den Pool der Wiederanrufer.
- **W1:** Ein Wiederanrufer trifft ein und stellt sich an die Warteschlange an bzw. wird sofort bedient.
- **W2:** Ein Wiederanrufer trifft ein, stellt fest dass es keinen freien Warteraum mehr gibt und verlässt das System endgültig. Oder aber: Ein Wiederanrufer trifft ein, legt trotz vorhandenem Warteraum sofort wieder auf und verlässt das System endgültig.
- **W3:** Ein Wiederanrufer trifft ein, stellt fest dass es keinen freien Warteraum mehr gibt und begibt sich in den Pool der Wiederanrufer. Oder aber: Ein Wiederanrufer trifft ein, legt trotz vorhandenem Warteraum sofort wieder auf und begibt sich in den Pool der Wiederanrufer.
- **W4:** Ein Wiederanrufer trifft ein, stellt fest dass er warten müsste und verlässt das System endgültig.
- **W5:** Ein Wiederanrufer trifft ein, stellt fest dass er warten müsste und begibt sich in den Pool der Wiederanrufer.
- **B:** Die Bedienung eines Kunden ist abgeschlossen.
- **A1:** Ein Kunde gibt das Warten auf und verlässt das System endgültig.
- **A2:** Ein Kunde gibt das Warten auf und begibt sich in den Pool der Wiederanrufer.

Zustandübergänge in der Q-Matrix



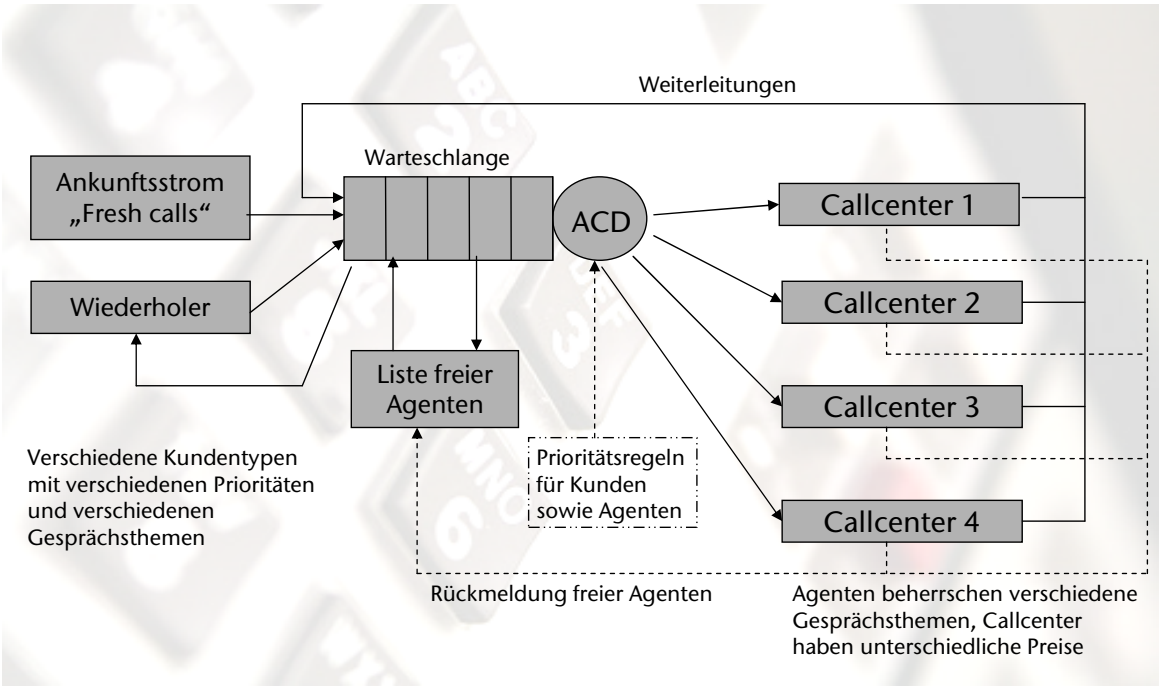
Numerisches Beispiel



Bei hoher Last unterschätzt das erweiterte Erlang-C-Modell die Anzahl der Kunden im System, da es die Wiederanrufer nicht berücksichtigt.

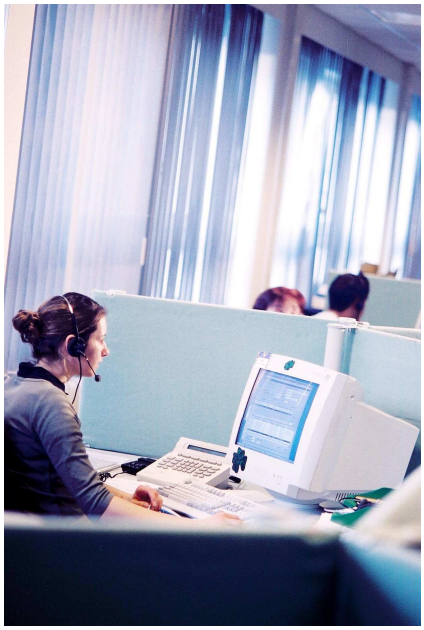
c=10, K=15, μ=10, ν=60, P_1=90%, P_2=80%, P_3=90%, α=0,5%, β=2%, δ=2

Aufbau eines realen Callcenter-Warteschlangenmodells



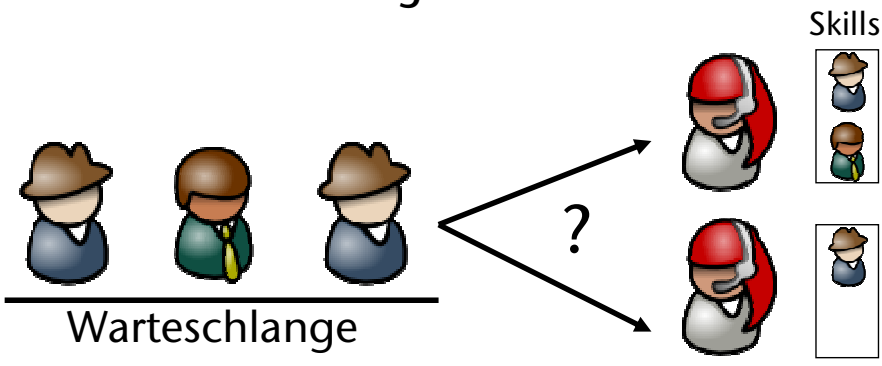
Problemstellungen in realen Callcentern

- Weiterleitungen
- Verschiedene Kundengruppen
- Verschiedene Agentengruppen mit verschiedenen Raten μ pro Kundengruppe (und evtl. generell verschiedenen Fähigkeiten)
- Schwankende Ankunfsraten und Agentenverfügbarkeiten (kein eingeschwungener Zustand mehr)
- Zuweisungsregeln der Kunden zu den Agenten (Single/Multi-Skill)
- Verschiedene Prioritäten der verschiedenen Kundengruppen



TU Clausthal

Single-Skill und Multi-Skill Agenten



- Single-Skill Agenten sind meist günstiger und auch schneller.
- Multi-Skill Agenten vermeiden Weiterleitungen.
- Ein hoher Anteil an Multi-Skill Agenten führt zu mehr Flexibilität und verringert das Risiko, dass Kunden warten müssen, obwohl Single-Skill Agenten (mit unpassenden Skill) bereit wären.
- Eine stets gültige optimale Zuweisungsstrategie existiert nicht.

Ereignisorientierte stochastische Simulation (1)

- Nachbildung des gesamten Warteschlangensystems durch Ereignisse.

- Beispiel für Ereignisse:

  - **Kunde trifft ein**: Erhöhe den Zähler „Kunden im System“ und prüfe, ob ein Bediener frei ist; wenn ja, löse das Ereignis „Bedienung beginnt“ aus.

  - **Bedienung beginnt**: Entferne den ersten Kunden aus der Warteschlange, bestimmte eine Bedienzeit gemäß der vorgegebenen Verteilung und führe zum Zeitpunkt t+Bedienzeit das Ereignis „Bedienung beendet“ aus.

  - **Bedienung beendet**: Verringere den Zähler „Kunden im System“ um eins. Ist mindestens ein Kunde in der Warteschlange, so führe das Ereignis „Bedienung beginnt“ aus.

TU Clausthal

Ereignisorientierte stochastische Simulation (2)

Ereignisliste

07:56 Kunde B kommt an
07:59 Kunde A verlässt das System (Gesprächsende)
07:59 Bediener geht in Nachbearbeitungszeit
08:00 Kunde C kommt an
08:01 Bediener meldet sich als frei
08:01 Kunde B geht zu Bediener
08:02 Kunde B gibt das Warten auf
08:06 Kunde C gibt das Warten auf
08:07 Kunde B verlässt das System (Gesprächsende)

Startsituation

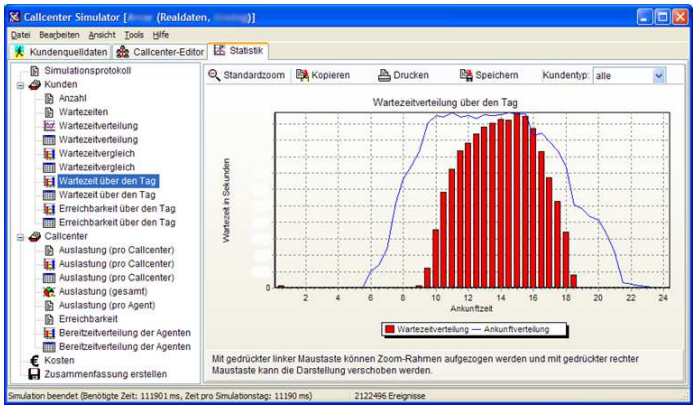
- Ein Bediener im System
- Kunde A ist befindet sich momentan in Bedienung

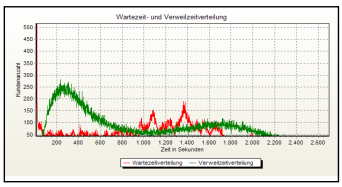
Algorithmus

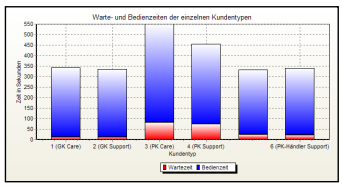
- Setze Uhrzeit auf Zeitpunkt des ersten Ereignisses.
- Arbeite das erste Ereignis ab und entferne es aus der Liste (der Ereignis kann dabei neue Ereignisse in die Liste eintragen; z.B. trägt das Ereignis „Gespräch beginnt“ das Ereignis „Gespräch Ende“ mit der Zeit t+Gesprächszeit ein.
- Setze Uhrzeit auf den Zeitpunkt des neuen ersten Ereignisses und setzt mit Punkt 2 fort. Sind alle Ereignisse abgearbeitet, so ist der Simulations-Tag zu Ende.

Callcenter Simulator

- Berücksichtigung aller relevanten Eingangsdaten eines realen Systems (bestehend aus mehreren Sub-Callcentern und mehreren Kunden- und Agentengruppen).
- Ausgabe von Kenngrößen wie Erreichbarkeit, Wartezeiten, Auslastungen usw. pro Kunden- bzw. Agentengruppe und gesamt sowie pro Halbstunden-Intervall und im Durchschnitt über den Tag.







Vielen Dank für Ihr Interesse.