Warteschlangenmodelle

Prof. Dr. Helmut Dietl

Lernziele

Nach dieser Veranstaltung sollten Sie,

- die strategische Bedeutung von Serverkapazitätsentscheidungen kennen
- die wichtigsten Warteschlangenmodelle zuordnen und anwenden können
- die wichtigsten Performancekriterien von Warteschlangensystemen berechnen können
- Kapazitätsentscheidungen auf der Basis von Warteschlangenmodellen treffen können

Methoden der Kapazitätsplanung in Service-systemen

- Warteschlangenmodelle
  - Schnelle Ergebnisse
  - Erfordern wenig Daten

- Simulationsmodelle
  - Können komplexe Sachverhalte berücksichtigen

- Lineare Programmierung
  - Zur Kapazitätsallokation über mehrere Einrichtungen/Standorte
  - Ermöglicht Integration der Reihenfolgeplanung und zusätzlicher Restriktionen

Herausforderungen der Kapazitätsplanung im Servicemanagement

- Nachfrageschwankungen führen zwangsläufig zu Auslastungsschwankungen
- Ungenützte Kapazität (keine Lager)
- Variable Kundenankunftsrate und variable Servicedauer
- Wartezeit und Kapazitätsauslastung sind Bestandteil der Servicequalität (Kunde als Koproduzent, z.B. verspätete oder volle Flugzeuge aber: ausverkaufte Konzerte, volle Diskotheken)
- Wegen der Nachfragevarianz wird die Kapazität i.d.R. in Input- (Bettenzahl) statt Outputgrößen (Gäste/Nacht) gemessen

Strategische Bedeutung von Kapazitätsent-scheidungen

- *Sumo-Strategie*: große Kapazität als Abschreckung potentieller Wettbewerber (z.B. 500-Betten-Luxushotel in Kleinstadt)

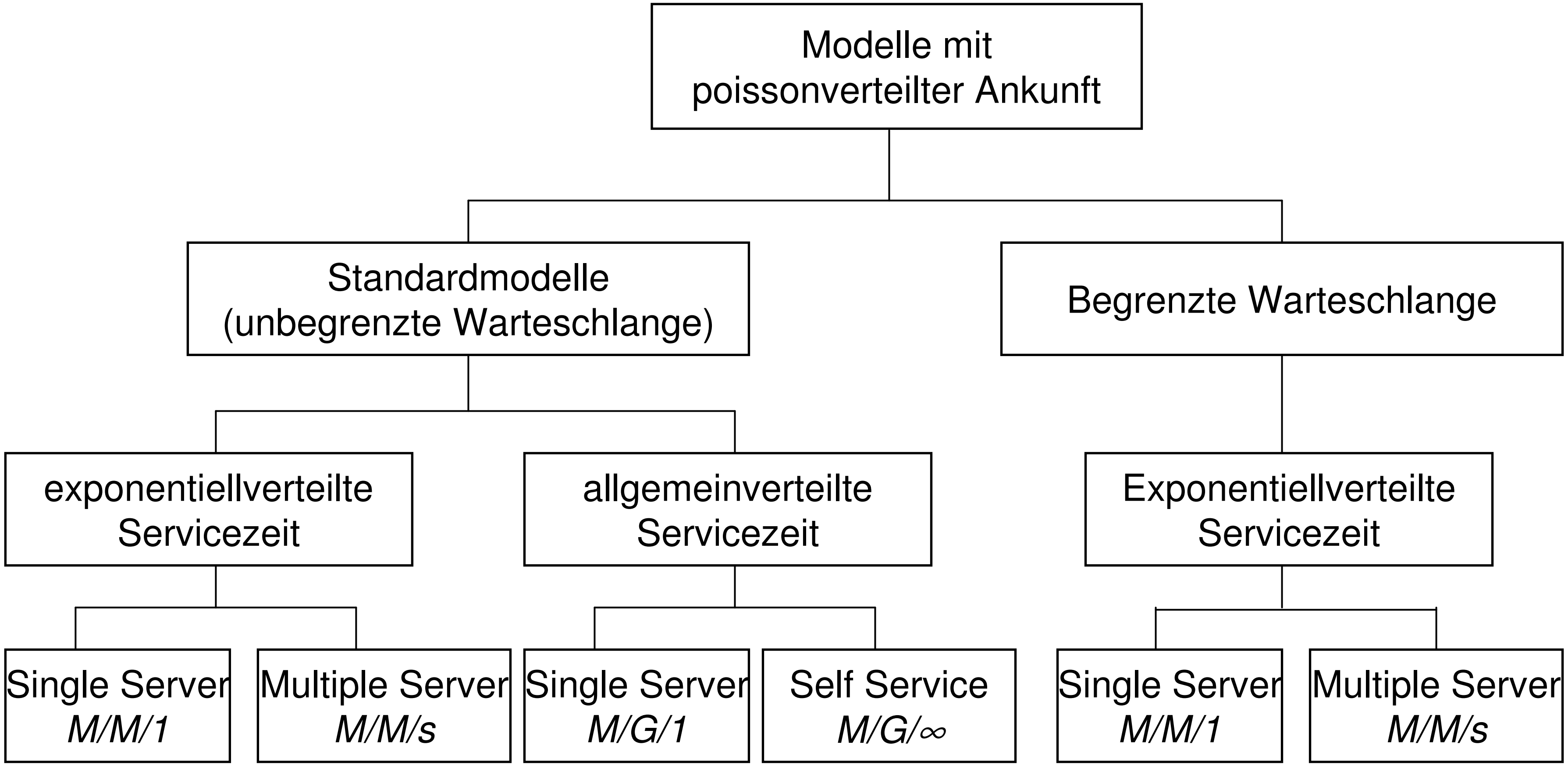
- *Judo-Strategie*: kleine Kapazität, um von starken Konkurrenten geduldet zu werden (z.B. Fluglinie mit 5 Maschinen)

- Unterkapazität generiert Nachfrage für Wettbewerber (überfülltes Restaurant)

- Über-/Unterkapazität sichert Kundenzufriedenheit (Beispiele: Überkapazität: Telekommunikation; Unterkapazität: Rockkonzert)

- Trade-off Optimierung: Umsatz/Kundenzufriedenheit versus Auslastungsgrad/Kundenzufriedenheit

Warteschlangenmodelle im Überblick



A/B/C Notation: *A* beschreibt die Verteilung der Zeitabstände zwischen 2 Ankünften, *B* beschreibt die Verteilung der Servicezeit und *s* (oder *c*) die Anzahl der Server.
M beschreibt die Exponentialverteilung, *G* irgendeine allgemeine Verteilung (z.B. Normalverteilung, Gleichverteilung, etc.)

M/M/1

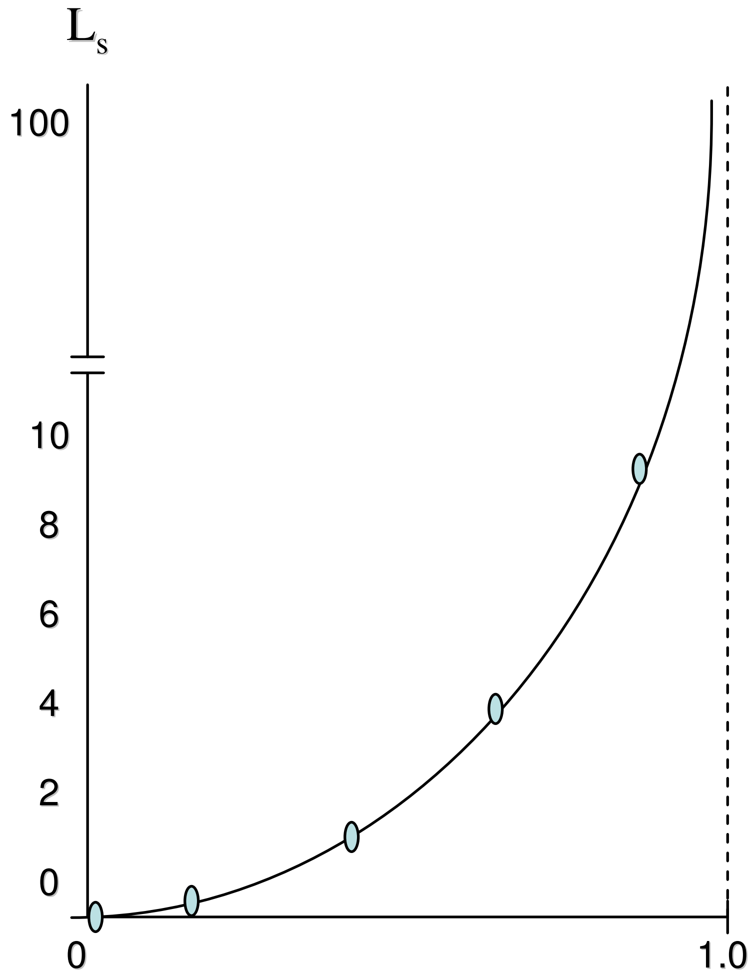
Durchschnittliche Anzahl von Kunden im System: $$Ls = \frac{\lambda}{\mu - \lambda}$$

Durchschnittliche Länge der Warteschlange: $$Lq = \frac{\rho\lambda}{\mu - \lambda}$$

Durchschnittliche Verweildauer im System: $$Ws = \frac{1}{\mu - \lambda}$$

Durchschnittliche Verweildauer in der Warteschlange: $$Wq = \frac{\rho}{\mu - \lambda}$$

Wie ändert sich die Länge der Warteschlange, wenn $\rho \to 0$?



$$\rho = \frac{\lambda}{\mu}$$

$$L_s = \frac{\rho}{1-\rho}$$

ρ	L_s
0	0
0.2	0.25
0.5	1
0.8	4
0.9	9
0.99	99

Beispiel 1: Eisverkäufer

- Pro Stunde kommen durchschnittlich 80 Kunden.
- Der Verkäufer benötigt je Kunde durchschnittlich 30 Sekunden.
- Ankunftsrate der Kunden ist poissonverteilt.
- Servicerate ist poissonverteilt.

Fragen zum Eisverkäufer-Beispiel

1. Wie hoch ist der durchschnittliche Auslastungsgrad des Eisverkäufers?

2. Wie lang ist die durchschnittliche Warteschlange vor dem Eisverkäufer?

3. Wie viele Kunden befinden sich durchschnittlich im „System“ (Warteschlange + Bedienung)?

4. Wie lange verweilt ein Kunde durchschnittlich in der Warteschlange (durchschnittliche Wartezeit)?

5. Wie lange verweilt ein Kunde durchschnittlich im „System“ (durchschnittliche Verweilzeit)?

Eisverkäufer-Beispiel

1. Durchschnittlicher Auslastungsgrad des Eisverkäufers

$$\lambda = 80 \text{ Kunden/Stunde}$$

$$\mu = \frac{1 \text{ Kunde}}{30 \text{ Sekunden}(1 \text{ Stunde} / 3600 \text{ Sekunden})} = 120 \text{ Kunden / Stunde}$$

$$\rho = \frac{\lambda}{\mu} = \frac{80 \text{ Kunden/Stunde}}{120 \text{ Kunden/Stunde}} = 0{,}67 = 67\%$$

Eisverkäufer-Beispiel

2. Durchschnittliche Länge der Warteschlange

$$Lq = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{80^2}{120(120 - 80)} = 1,33$$

3. Durchschnittliche Anzahl der Kunden im System

$$Ls = \frac{\lambda}{\mu - \lambda} = \frac{80}{120 - 80} = 2$$

Eisverkäufer-Beispiel

4. Durchschnittliche Wartezeit

$$Wq = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{80}{120(120 - 80)} = \frac{1}{60}\,\text{Stunde} = 1\,\text{Minute}$$

5. Durchschnittliche Verweilzeit

$$Ws = \frac{1}{\mu - \lambda} = \frac{1}{120 - 80} = \frac{1}{40}\,\text{Stunde} = 1{,}5\,\text{Minuten}$$

M/M/1-Modell mit begrenzter Warteschlange

- Eisverkäuferbeispiel:
  - Eisverkäufer möchte einen McIce Drive in Kiosk betreiben.
  - Wie viele Autos müssen in der Drive-in-Schlange mindestens Platz haben, damit mit weniger als 10%iger Wahrscheinlichkeit keine Autos auf der Strasse warten müssen?
  - Ansonsten wie zuvor
  - Lösung: $P(n \geq k) = \rho^k$ $P(n \geq 5) = \rho^5 = 0.13$

$$P(n \geq 6) = \rho^6 = 0.09$$

M/G/1-Modell

$$Lq = \frac{\rho^2 + \lambda^2 \sigma^2}{2(1-\rho)}$$

1. Für Exponentialverteilung gilt: $\sigma^2 = \frac{1}{\mu^2}$ => $Lq = \frac{\rho^2 + \lambda^2 / \mu^2}{2(1-\rho)} = \frac{2\rho^2}{2(1-\rho)} = \frac{\rho^2}{(1-\rho)}$

2. Bei konstanter Servicezeit gilt: $\sigma^2 = 0$ => $Lq = \frac{\rho^2}{2(1-\rho)}$

3. Hieraus folgt, dass die durchschnittliche Länge der Warteschlange (L_q) jeweils zur Hälfte durch die Varianz der Ankünfte sowie die Varianz der Servicezeit erklärt wird.

Standard-M/M/s-Modell

- Voraussetzungen:
  - Wie bei Standard M/M/1-Modell (u.a. eine unbegrenzte Warteschlange und FCFS)
  - Servicerate der Server ist unabhängig und identisch verteilt
  - $\lambda < s\mu$ bzw. $\lambda/\mu = \rho < s$

mms.xls M/M/s Queueing Formula Spreadsheet

Inputs:

lambda	80
mu	120

Definitions of terms:

lambda = arrival rate
mu = service rate
s = number of servers
Lq = average number in the queue
Ls = average number in the system
Wq = average wait in the queue
Ws = average wait in the system
P(0) = probability of zero customers in the system
P(delay) = probability that an arriving customer has to wait

Outputs:

s	Lq	Ls	Wq	Ws	P(0)	P(delay)	Utilization
0							
1	1,3333	2,0000	0,0167	0,0250	0,3333	0,6667	0,6667
2	0,0833	0,7500	0,0010	0,0094	0,5000	0,1667	0,3333
3	0,0093	0,6760	0,0001	0,0084	0,5122	0,0325	0,2222
4	0,0010	0,6677	0,0000	0,0083	0,5133	0,0051	0,1667
5	0,0001	0,6668	0,0000	0,0083	0,5134	0,0007	0,1333
6	0,0000	0,6667	0,0000	0,0083	0,5134	0,0001	0,1111
7	0,0000	0,6667	0,0000	0,0083	0,5134	0,0000	0,0952
8	0,0000	0,6667	0,0000	0,0083	0,5134	0,0000	0,0833
9	0,0000	0,6667	0,0000	0,0083	0,5134	0,0000	0,0741
10	0,0000	0,6667	0,0000	0,0083	0,5134	0,0000	0,0667

Beispiel: Fahrkartenautomat

- Pro Stunde kommen durchschnittlich 20 Kunden
- Servicemanager hat die Wahl zwischen
  - Einem modernen Hochleistungsautomat (bedient durchschnittlich 2 Kunden pro Minute)
  - Zwei alten Automaten (bedienen jeweils durchschnittlich 1 Kunden pro Minute)

Fahrkartenbeispiel: 1 Hochleistungsautomat

mms.xls M/M/s Queueing Formula Spreadsheet

Inputs:

lambda	20
mu	120

Definitions of terms:

lambda = arrival rate
mu = service rate
s = number of servers
Lq = average number in the queue
Ls = average number in the system
Wq = average wait in the queue
Ws = average wait in the system
P(0) = probability of zero customers in the system
P(delay) = probability that an arriving customer has to wait

Outputs:

s	Lq	Ls	Wq	Ws	P(0)	P(delay)	Utilization
0							
1	0,0333	0,2000	0,0017	0,0100	0,8333	0,1667	0,1667
2	0,0012	0,1678	0,0001	0,0084	0,8462	0,0128	0,0833

Fahrkartenbeispiel: 2 Altautomaten

mms.xls M/M/s Queueing Formula Spreadsheet

Inputs:

lambda	20
mu	60

Definitions of terms:

lambda = arrival rate
mu = service rate
s = number of servers
Lq = average number in the queue
Ls = average number in the system
Wq = average wait in the queue
Ws = average wait in the system
P(0) = probability of zero customers in the system
P(delay) = probability that an arriving customer has to wait

Outputs:

s	Lq	Ls	Wq	Ws	P(0)	P(delay)	Utilization
0							
1	0,1667	0,5000	0,0083	0,0250	0,6667	0,3333	0,3333
2	0,0095	0,3429	0,0005	0,0171	0,7143	0,0476	0,1667
3	0,0006	0,3340	0,0000	0,0167	0,7164	0,0050	0,1111

Trade-offs

- **2 Altautomaten**
  Lq = 0.0095
  Ls = 0.3429
  Wq = 0.0005
  Ws = 0.0171
  P(0) = 71%
  P(Delay) = 4.8%
  Auslastungsgrad = 16.7%

- **1 Hochleistungsautomat**
  Lq = 0.0333
  Ls = 0.2
  Wq = 0.0017
  Ws = 0.01
  P(0) = 83%
  P(Delay) = 16.7%
  Auslastungsgrad = 16.7%

Server-Pooling

- Prinzip: Eine statt mehrere Warteschlangen
- Bessere Auslastung der Server
- Beispiele: Postschalter, Sekretärinnenpool
- Nachteil: „Lange“ Warteschlange schreckt evtl. Kunden ab
- Trade off zwischen Transport- und Wartekosten bei Pooling über mehrere Standorte (1 zentraler Serverpool vs. mehrere dezentrale Server)

M/M/s-Modell mit begrenzter Warteschlange

- Analog zu M/M/1-Modell mit begrenzter Warteschlange
- N = Maximale Kundenzahl im System > s
- Neu ankommender Kunde wird zurückgewiesen, wenn mehr als N-s Kunden warten oder mehr als N Kunden im System sind
- Sonderfall: N - s = 0 (Keine Wartemöglichkeit)
- Beispiel: Parkplatz (jeder Parkplatz ist ein Server)

M/G/∞-Modell

- Bei diesem Modell muss kein Kunde warten, da es unendliche viele Server gibt
- Beispiel: Selbstbedienung
- Die Anzahl der Kunden im System ist poissonverteilt gemäß

$$P_n = \frac{e^{-\rho}}{n!} \rho^n$$

- Es gilt

$$Ls = \rho$$

Beispiel: Supermarkt

- Supermärkte können als 2 sukzessive Warteschlangensysteme modelliert werden
- System 1: Selbstbedienung im Laden (M/G/ ∞)
- System 2: Kasse (falls eine Schlange mit mehreren Kassen: M/M/s)

Kostenminimierung

Gesamtkosten pro Stunde = Wartekosten pro Stunde + Servicekosten pro Stunde

$$TC = C_w \lambda W_s + C_s C = C_w L_s + C_s C$$

C_w = Opportunitätskosten/Stunde eines Kunden

λ = durchschnittliche Ankunftsrate

W_s = durchschnittliche Verweilzeit im System

C_s = Serverkosten/Stunde

C = Serverzahl

Achtung: Gilt nur für Systeme mit $C > \rho = \lambda/\mu$

Beispiel: Workstation Miete

- Ein Ingenieurbüro plant Workstations für die Durchführung von Statikanalysen anzumieten
- Durchschnittlich werden pro Stunde 8 Statikanalysen durchgeführt (poissonverteilt)
- Eine Statikanalyse dauert durchschnittlich 15 Minuten (exponentialverteilt)
- Die Mietkosten betragen je Workstation 10€/h
- Der Stundenlohn eines Ingenieurs beträgt 30€/h

Workstation-Beispiel

Berechnung: M/M/s-Modell mit ρ = 8/4 = 2

C	L_q	$C_w\ L_q$	C_sC	TC
3	0.88	€26.4	€30	€56.4
4	0.17	5.1	40	45.1
5	0.04	1.2	50	51.2
6	0.01	0.3	60	60.3

Achtung: Hier wird mit Lq anstatt Ls gerechnet, da die Ingenieure an der Workstation bereits produktiv sind!