3 Analytische Techniken für diskrete Systeme

Bisher wurden ereignisdiskrete Systeme analysiert, indem ihr Verhalten im Rechner „nachgespielt“ wurde

⇒ Trajektorie des unterliegenden stochastischen Prozesses konnte beobachtet und analysiert werden

- Ergebnisse sind statistischen Schwankungen unterworfen
- Resultate sind nur in Form von Konfidenzintervallen ermittelbar
  (d.h. mit vorgegebener Wahrscheinlichkeit liegt der wahre Wert im geschätzten Intervall)
- Je nach Typ des zu ermittelnden Resultats sind sehr viele/lange Simulationsläufe notwendig (⇒ hoher Aufwand)
- Gewisse Resultate können per Simulation überhaupt nicht zuverlässig ermittelt werden (z.B. kleine Wahrscheinlichkeiten)

Insgesamt gilt:

Simulation sollte immer des letzte Mittel sein, falls andere effizientere Techniken versagen oder nicht anwendbar sind!!

Welche anderen Techniken gibt es?

Naheliegend ist die numerische/analytische Analyse des (i.d.R. vereinfachten) stochastischen Prozesses

- Numerische/analytische Berechnungen liefern (bis auf numerische Rundungsfehler) exakte Resultate auch für kleine Wahrscheinlichkeiten etc.

Warum sind diese Techniken effizienter?

- Effizientere Berechnungen
- Abstraktere Modelle
- Geringerer Aufwand der Modellerstellung
- Geringerer Aufwand der Datenerhebung und Datenmodellierung

Existierende Verfahren (Auswahl):

- (Numerische) Analyse von Markov-Prozessen
  - d.h. numerische Analyse eines linearen Gleichungssystems oder Differentialgleichungssystems
- (Analytische) Resultate für spezielle Wartesysteme
  - geschlossene Formeln für zahlreiche Leistungsmaße sind bekannt
- (Quasi analytische) Verfahren für bestimmte Klassen von Warteschlangennetzen
  - effiziente Analysealgorithmen existieren für sogenannte Produktformnetze
- Zahlreiche Approximationsmethoden für allgemeinere Warteschlangennetze
  - oft basierend auf Heuristiken aber mit passabler Ergebnisgenauigkeit bei geringem Analyseaufwand

Insgesamt ein weites Feld, welches mehr als eine Vorlesung füllen kann

Ziele:

- Potenzial vorhandener analytisch/numerischer Analysetechniken bewerten können
- Kennen lernen einiger grundlegender Gesetzte der analytischen Leistungsanalyse
- Kennen lernen der analytischen Berechnung von Leistungsgrößen einiger elementarer Wartesysteme
- Übersicht über eine Klasse offener Warteschlangennetze mit Produktformlösung erhalten

Gliederung:

3.1 Einfache Stationen

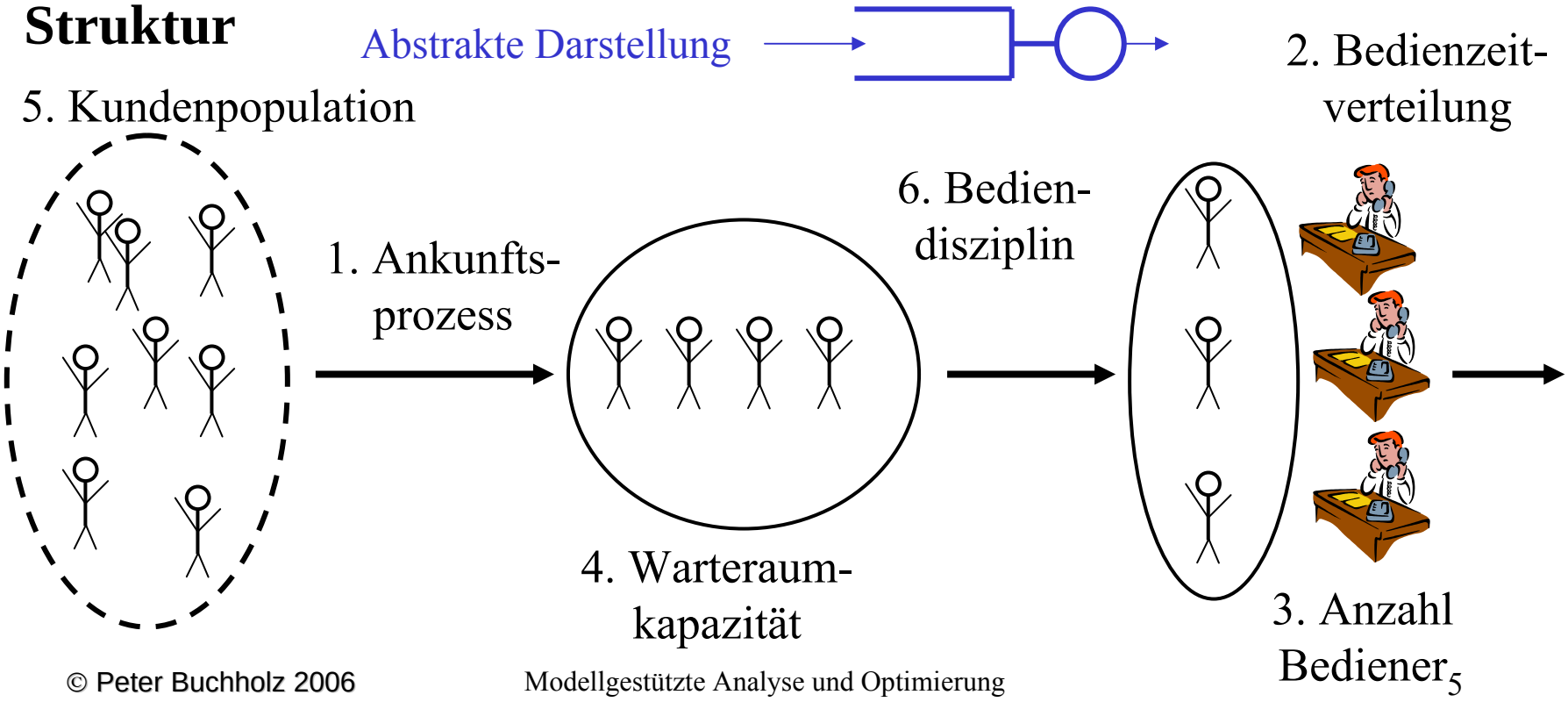
3.2 Offene Warteschlangennetze

3.1 Einfache Stationen

Basissystem als Generalisierung des mehrfach benutzten Schalterbeispiels

Bekannt unter mehreren Bezeichnungen Warteschlangen, Bediensysteme, Wartesysteme, Stationen, ... (+ diverse engl. Bez.)

Struktur



1. Ankunftsprozess: Stochastischer Prozess zur Beschreibung der Kundenankünfte,
2. Bedienzeitverteilung: Verteilung der ZV zur Definition der Bedienzeiten

Typische Beispiele (mit Abkürzungen) werden für Bedienzeiten und Zwischenankunftszeiten verwendet

- Exponential-Verteilung (M)
- Erlang k-Verteilung (E_k)
- Hyperexponential-Verteilung mit k Phasen (H_k)
- Deterministische-Verteilung (D)
- Allgemeine-Verteilung (G) zus. unabhängig (GI)
- ....

Bisherige Definition umfasst einzelne Ankünfte und einzelne Bedienungen, natürliche Erweiterung auf Gruppen-Ankünfte und Gruppen-Bedienung:

- Symbolische Darstellung A^B, wobei
  - $A \in \{M, E_k, H_k, D, G, GI, ...\}$ Verteilung der Zwischenankunftszeit
  - $B \in \{M, Geo, D, G, GI, ...\}$ diskrete Verteilung der Gruppengröße
    M bedeutet hier Poisson-Prozess, Geo geometrische Verteilung

3. Anzahl Bediener
   alle Bediener werden als identisch angenommen
   falls die Anzahl Bediener unendlich ist, spricht man auch von einer Verzögerungsstation
4. Größe des Warteraums
   es wird davon ausgegangen, dass Kunden bis zum Ende ihrer Bedienung im Warteraum bleiben (Größe des Warteraums ≥ Anzahl Bediener)
   Kunden, die eintreffen, wenn der Warteraum vollständig gefüllt ist, werden abgewiesen
   a. falls Anzahl Bediener = Größe Warteraum,
      spricht man auch von einem Verlustsystem
   b. falls Anzahl Bediener < Größe Warteraum < ∞,
      spricht man auch von einem Warte-/Verlustsystem
5. Kundenpopulation im System
   potenzielle Population, die das System nutzen könnte
   neben der maximalen Kundenzahl hängt die Ankunftsintensität von der Kundenpopulation ab
6. Bediendisziplin (Auswahl der Kunden zur Bedienung)
   Beispiele: FCFS (First Come First Served), Random, LCFS (Last Come First Served), PS (processor sharing), SPT (shortest processing time first), ...

Spezifikation von Stationen über Kendall-Notation:
A/B/c/N/K/SD

- A Zwischenankunftszeit (ZV)
- B Bedienzeit (ZV)
- c Anzahl Bediener
- N Kapazität des Warteraums (Voreinstellung ∞)
- K Gesamtpopulation (Voreinstellung ∞)
- SD Bedienstrategie (Voreinstellung FCFS)

Beispiele:

M/M/1 exponentiell-verteilte Zwischenankunfts- und Bedienzeiten, ein Bediener, unendliche Kapazität des Warteraums und unendliche Population

M/GI/2/5/10/RANDOM exponentiell-verteilte Zwischenankunftszeiten, unabhängig allgemein-verteilte Bedienzeiten, 2 Bediener, Warteraum mit Kapazität 5, Gesamtpopulation 10 und zufällige Auswahl der Kunden

Unterschiedliche Interpretationen von Stationen in verschiedenen Anwendungsgebieten:

System	**Kunden**	**Bediener**
Bank	Kunden	Bankkaufmann/-frau
Werkstatt	defekte Maschinen	Handwerker/-in
Krankenhaus	Patienten	Arzt/Ärztin
Lager	Paletten	Gabelstapler
Straße	Autos	Ampel
Rechner	Prozesse	CPU
Datenbank	Transaktionen	DB-Server
Fertigungslinie	Werkstück	Maschine

Vielfältige Möglichkeiten der Interpretation!

Analyse auch hier wieder stationär oder transient.
Wir beschränken uns auf die stationäre Analyse und nehmen an, dass das System einen stationären Zustand erreicht.

Parameter des Systems:

- $\lambda = E(A)^{-1}$ Ankunftsrate
- $\mu = E(B)^{-1}$ Bedienrate

Ergebnisgrößen (ZV wir betrachten nur die Erwartungswerte):

- X Durchsatz (= Kunden pro Zeiteinheit, die bedient werden)
- Q Population (= Kunden im System)
- R Verweilzeit eines Kunden im System
- U Auslastung (= Zeitanteil, den ein Bediener belegt ist)

Zusätzlich

- p(i) Wahrscheinlichkeit, dass sich i Aufträge im System befinden

Einige Zusammenhänge:

$E(U) = 1 - p(0)$ (bei einem Bediener)

$$E(Q) = \sum_{i=1}^{\infty} i \cdot p(i)$$

$$E(X) = \sum_{i=1}^{m-1} i \cdot p(i) \cdot \mu + \sum_{i=m}^{\infty} m \cdot p(i) \cdot \mu$$

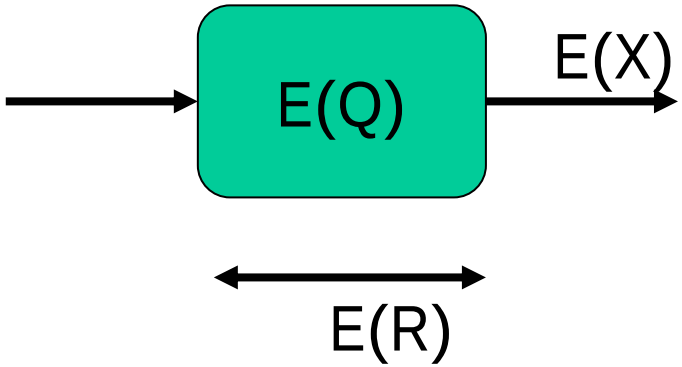
Bei Einbediener-Systemen $E(X) = (1 - p(0)) \cdot \mu$

und $E(U) = \rho = \lambda / \mu$ falls $\lambda < \mu$

Operationale Gesetzte gelten für allgemeine Systeme unabhängig von der Ankunftszeit, Bedienzeit oder Bedienstrategie!!

Gesetz von Little (gilt allgemein für Systeme im Gleichgewicht)

System als „Black Box“, an der Kunden ankommen, bearbeitet werden und weggehen.
Das interne System kann einfach oder komplex sein.



Es gilt $E(Q) = E(X) \cdot E(R)$

Beobachtete Größen:

- Es sind im Mittel E(Q) Kunden im System.
- Ein Kunde verweilt im Mittel E(R) Zeiteinheiten im System.
- Das System hat einen mittleren Durchsatz von E(X).

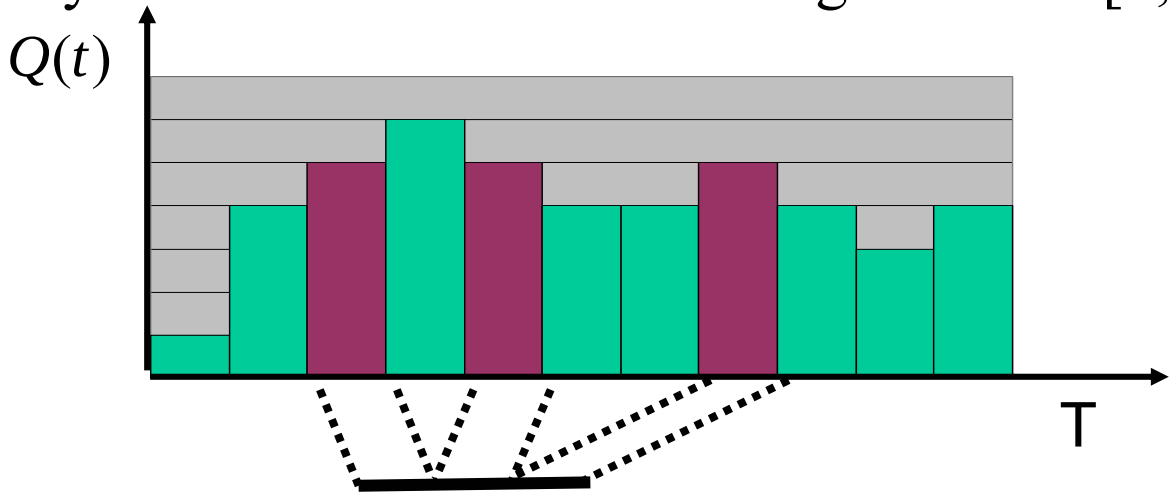
Beispiel:

- LAN Router hat Latenzzeit von 200 μs/ Paket bei einer Übertragungsrate von 30000 Paketen/s.
  Mittlere Paketzahl im Router ?

$$E(Q) = 30000 \cdot 0.0002 = 6$$

Herleitung des Gesetztes: System wir im Intervall [0,T] beobachtet

- Systemzustand im Beobachtungsintervall [0,T]



- f_k prozentualer Anteil am Intervall [0,T] zu dem sich k Kunden im System befinden
- r_k Zeitdauer, zu der sich k Kunden im System befinden, damit gilt $r_k = f_k \cdot T$
- Weiterhin gilt $\overline{Q} = \sum_k k \cdot f_k = \sum_k k \cdot r_k / T$
- C_0 Gesamtzahl Kunden, die das System im Beobachtungsintervall verlassen haben

$$\overline{Q} = \frac{C_0}{T} \frac{\sum_k k \cdot r_k}{C_0} = \overline{X} \cdot \overline{R} \quad (\overline{Q}, \overline{R}, \overline{X} \text{ Mittelwerte})$$

Das M/M/1-System

Bisher vorgestellte Gesetze gelten zwar allgemein, erlauben aber noch keine Berechnung der Leistungsgrößen auf Basis der Systemparameter

$\Rightarrow$ Untersuchung einer eingeschränkten Modellklasse

Einfachste Variante: Das schon bekannte M/M/1-System

- Ankünfte
  - durch einen Poisson-Strom realisiert
    Wahrscheinlichkeit für k Ankünfte im Intervall [0,t):
    $P[Y=k] = e^{-\lambda t}(\lambda t)^k / k!$
  - Zwischenankunftszeiten negativ exponentiell verteilt
    $FA(t)= 1 - e^{-\lambda t}$
- Bedienungen
  - Bedienzeiten negativ exponentiell verteilt
    $FB(t)= 1 - e^{-\mu t}$

Weitere Charakteristika des Systems

- 1 Bediener
- unendliche Anzahl von Warteplätzen
- unendliche Population (d.h. konstante Ankunftsrate)
- Bedienstrategie FCFS (erst einmal, später auch andere Strategien)

Simulation des Systems $\Rightarrow$ einfacher Schalter

Realisierung dort

- Zustand beschrieben durch ganze Zahl (= Kunden im System)
- Zwei Ereignisse
  - Ankunft: belege Schalter; falls leer, erhöhe Kundenzahl um1
  - Bedienende: starte Bedienung nächster Kunde (falls vorhanden), erniedrige Kundenzahl um 1

$\Rightarrow$ einfache Realisierung, problemlose Simulation?

Vorgehen bei der Simulation:

Transiente Simulation (d.h. Ermittlung der Leistungsmaße zum Zeitpunkt t ausgehen vom bekannten Zustand zum Zeitpunkt 0):

- Wiederholte Simulation von 0 bis t (n Replikationen)
- Resultate unabhängig bei geeigneter Wahl der Saaten des ZZ-Generators
- Resultate normalverteilt falls n *genügend groß*

Stationäre Simulation (d.h. Ermittlung der Leistungsmaße für das „typische“ Systemverhalten):

- Wiederholte Simulation von 0 bis t (n Replikationen)
  Resultate stationär fall t *genügend groß*

oder

- ein langer Simulationslauf und Bildung von Batches der Größe b (insges. n Batches), Resultate unabhängig falls b *genügend groß*
- Resultate normalverteilt falls n *genügend groß*

In beiden Fällen Konfidenzintervalle via Normal- oder t-Verteilung

Beispiel:

Bedienrate $\mu = 1.0$, Ankunftsrate $\lambda = 0.05, 0.5, 0.95, 0.99$

Ziel Schätzung der mittlere Population E(Q)

Ergebnisse aus 100 Replikationen der Experimente

Population zum Zeitpunkt t = 100

λ	min	max	$\hat{Q}$	$\hat{S}(Q)$
0.05	0	2	0.100	0.333
0.5	0	8	1.060	1.693
0.95	0	31	7.510	6.711
0.99	0	43	10.16	8.178

Population zum Zeitpunkt t = 1000

λ	min	max	$\hat{Q}$	$\hat{S}(Q)$
0.05	0	2	0.070	0.293
0.5	0	8	0.970	1.410
0.95	0	67	18.07	16.89
0.99	0	116	31.70	26.51

Erste Interpretation:

- t = 100 sicher nicht ausreichend
- t = 1000 scheint ausreichend für $\lambda = 0.05$ und 0.5
  - Konfidenzintervalle nach Tsch. mit $\alpha = 0.1$: [0.0, 0.163] bzw. [0.524, 1.416]
  - Konfidenzintervalle unter Annahme normalverteilter Schätzer mit $\alpha = 0.1$: [0.022, 0.118] bzw. [0.737, 1.135]

Population zum Zeitpunkt t = 2000

λ	min	max	$\hat{Q}$	$\hat{S}(Q)$
0.95	0	84	21.71	20.47
0.99	0	177	41.16	37.43

- Immer noch deutlicher Anstieg des Mittelwertschätzers (transiente Phase dauert an !?)
- Berechnung der Konfidenzintervalle liefert
  - nach Tsch. mit $\alpha = 0.1$: [15.25, 28.19] bzw. [25.59, 59.0]
  - Normalverteilung mit $\alpha = 0.1$: [18.34, 25.10] bzw. [40.98, 53.34]
- In allen Fällen (zu) breite Konfidenzintervalle trotz moderatem α (Zur Erinnerung: Mit Wahrscheinlichkeit 10% kann der wahre Wert außerhalb des Konfidenzintervalls liegen, wenn die Voraussetzungen zur Berechnung stimmen.)

Beobachtungen aus dem kleinen Beispiel:

- Bei hoher Auslastung Unsicherheit, ob transiente Phsse beendet
- Damit lange Läufe bei hoher Auslastung und viele Replikationen für schmale Konfidenzintervalle
  (Konfidenzintervallbreite < 10% des ermittelten Mittelwertes erfordert im Beispiel für $\lambda = 0.5$ und $\alpha = 0.1$ ca. 1800 Replikationen)

Also schon für dieses einfache Beispiel zeigen sich Probleme der Simulation durch den hohen Aufwand und die Unsicherheit bei der Resultatermittlung

Dies gilt insbesondere bei

- hohen Auslastungen
- seltenen Ereignissen

Zur Verdeutlichung ein (sehr einfaches) Beispiel:
Tritt ein Fehler in einem vorgegebenem Intervall auf?

z.B. Sonde auf dem Weg zu einem anderen Planeten, Bearbeitung eines Werkstücks auf einer Maschine, ...

- Analyse durch mehrmalige Simulation des Ablaufs bis zum Intervallende, pro Replikation eine Realisierung der Resultat-ZV Y
- Resultat Y: 0 = ok, 1 = Fehler (üblicherweise p[Y=1] << 1.0)
- Ziel: Bestimmung von p(1) mit einer Genauigkeit von plus/minus 10% mit 99% Wahrscheinlichkeit
- Fragestellung: Wie viele Abläufe müssen simuliert werden?

Resultatschätzer $\tilde{p} = 1/n \sum_{i=1}^{n} y_i$ wobei y_i Ausgang der i-ten Beobachtung (0 oder 1)

Es gilt $E(\tilde{p}) = p$ und $\sigma^2(\tilde{p}) = p(1-p)/n$

Konfidenzintervall: $\hat{p} \pm 2.576 \cdot \hat{S}/\sqrt{n}$ mit $\hat{S}^2 = 1/(n-1) \sum_{i=1}^{n} (y_i - \hat{p})^2$

Da $E(\tilde{S}^2) = n \cdot \sigma^2(\tilde{p}) = p(1-p)$

ist der Erwartungswert der halben Breite des Konfidenzintervalls gleich

$2.576 \cdot \sqrt{p(1-p)/n}$

Es soll damit gelten $2.576 \cdot \sqrt{p(1-p)/n} \leq 0.1 \cdot \hat{p}$

Da $\lim_{n\to\infty} \hat{p} = p$ muss gelten $n \approx 100 \cdot 2.576^2 \cdot (1-p)/p = 663.58 \cdot (1-p)/p$

Je kleiner p ist, desto größer muss n gewählt werden!

Beispiele:

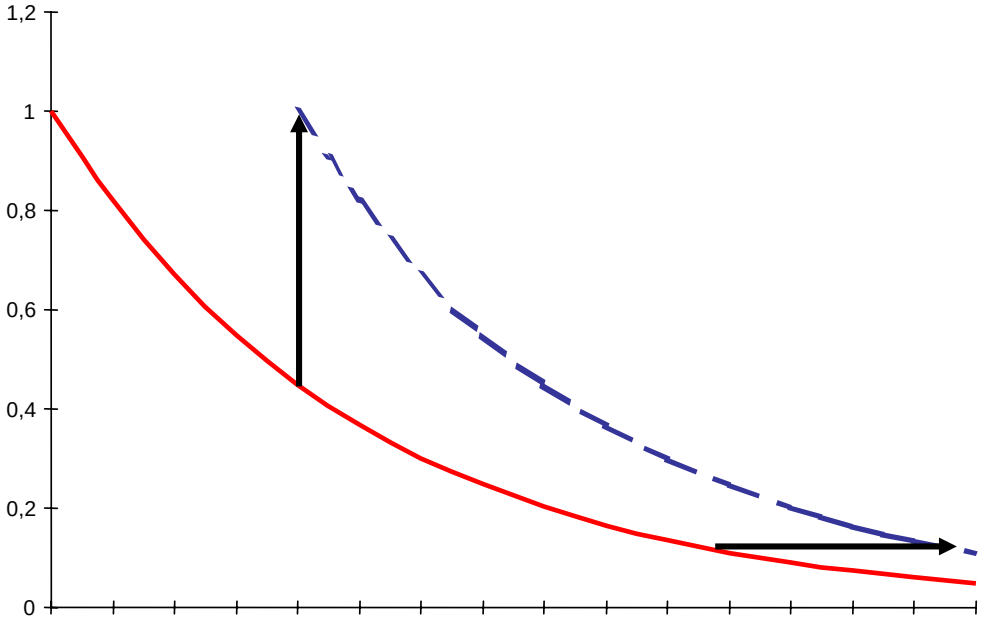
- $p = 0.1 \Rightarrow n \approx 5973$
- $p = 0.001 \Rightarrow n \approx 6{,}62{,}917$
- $p = 10^{-6} \Rightarrow n \approx 6.63 \cdot 10^8$
- $p = 10^{-8} \Rightarrow n \approx 6.63 \cdot 10^{10}$

Selbst für einfache Modelle auf schnellen Rechnern kaum durchführbar!

Eigenschaften der Exponential-Verteilung:

- Gedächtnislosigkeit (als einzige kontinuierliche Verteilung)

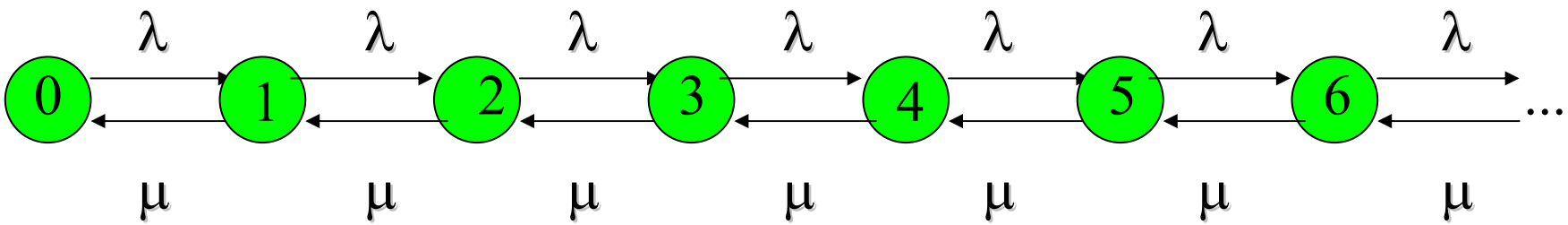
Restzeit bleibt gleich verteilt
$F(t+x|x) = F(t)$
damit ist die Zeit seit dem Beginn der Laufzeit irrelevant für die Vorhersage der Zukunft



- Additivität: Das Minimum zweier Exponential-Verteilungen mit Raten λ und μ ist exponentiell-verteilt mit Rate $(\lambda + \mu)$
  (Offensichtlich kann dies auf eine beliebige Anzahl von Exponential-Verteilungen erweitert werden)

Zustands-/Transitionsdiagramm für M/M/1:

- Zustände definiert durch die Anzahl Kunden im System
- Transitionen durch Kunden Ankünfte oder Abgänge (gewichtet mit der zugehörigen Rate)



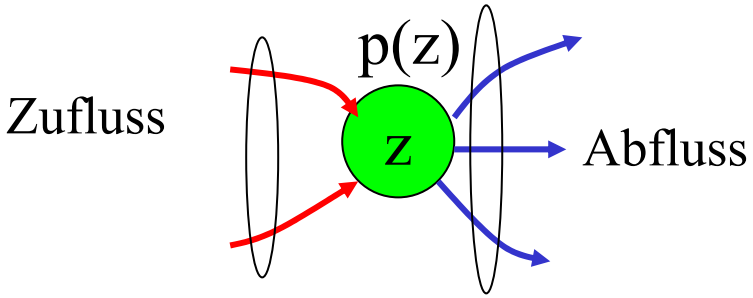
Verhalten des Systems im Zustand

- n=0 die Zeit bis zur Ankunft des nächsten Kunden ist exponentiell-verteilt mit Rate λ unabhängig von der bisherigen Verweilzeit im Zustand (Gedächtnislosigkeit)
- n>0 die Zeit bis zur nächsten Zustandsänderung ist exponentiell-verteilt mit Rate μ+λ (Additivität)
  - mit W. μ/(μ+λ) geht ein Kunde ab, mit W. λ/(μ+λ) kommt ein Kunde an

Wahrscheinlichkeitsfluss: W. im Zustand zu sein · Transitionsrate

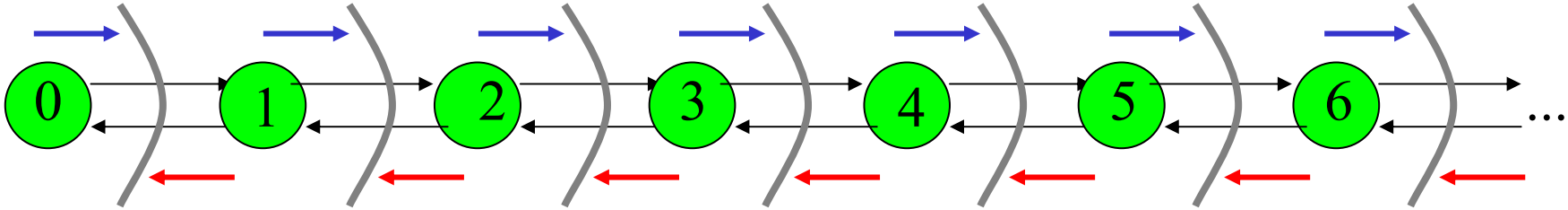
Im stationären Zustand ändert sich die Wahrscheinlichkeitsverteilung nicht ⇒

Fluss aus einem Zustand = Fluss in den Zustand



⇒ Berechnung der stationären Wahrscheinlichkeiten lässt sich auf Lösung eines linearen Gleichungssystems reduzieren (später mehr dazu)

Auf Grund der speziellen Struktur des M/M/1-Systems (Geburts-/Todes-Prozess):



Flussgleichgewicht über jeden Schnitt!

Resultierendes Gleichungssystem

$$p(i)\cdot\lambda = p(i+1)\cdot\mu \Rightarrow p(i) = p(0)\cdot\left(\frac{\lambda}{\mu}\right)^i = p(0)\cdot\rho^i$$

- Falls wir p(0) kennen, können wir alle p(i) berechnen
- Aus der Kenntnis der p(i) lassen sich die Leistungsmaße des Systems berechnen

Überlegungen zur Berechnung von p(0):
p(i) (i = 0, ..., ∞) definieren eine Wahrscheinlichkeitsverteilung
$\Rightarrow$ Summe der p(i) muss 1.0 ergeben

Damit gilt: $$\sum_{i=0}^{\infty} p(i) = p(0)\cdot\sum_{i=0}^{\infty}\rho^i \Rightarrow p(0) = \frac{1.0}{\sum_{i=0}^{\infty}\rho^i}$$

Voraussetzung Reihensumme endlich!

Geometrische Reihe: $$\sum_{i=0}^{\infty}\rho^i = (1-\rho)^{-1} \quad \text{falls } \rho < 1$$

Falls $\lambda > \mu$:

- Kommen im Mittel mehr Kunden an, als bedient werden können,
- damit wächst die unerledigte Arbeit (= Anzahl wartender Kunden)
- damit existiert kein stationärer Zustand

Was passiert bei $\lambda = \mu$?
Mathematik zeigt Verhalten wie bei $\lambda > \mu$.
Intuitive Erklärung dafür:

- Im Mittel kommt genau soviel Arbeit hinzu, wie der Bediener leisten kann
- Durch die stochastischen Schwankungen wird der Bediener von Zeit zu Zeit nicht belegt sein, dabei geht Bedienkapazität verloren
- Verloren gegangene Bedienkapazität fehlt später, dadurch wächst die Warteschlange und es existiert keine stationäre Verteilung

Für $\lambda \geq \mu$ kann das System nur für einen endlichen Zeithorizont analysiert werden

Berechnung der Leistungsmaße:

- Auslastung: $1 - p(0) = \rho$ und Durchsatz E(X) = λ
- Mittlere Kundenzahl:

$$
\begin{aligned}
E(Q) &= \sum_{i=0}^{\infty} i \cdot p(i) \\
&= (1-\rho) \cdot \sum_{i=0}^{\infty} i \cdot \rho^i \\
&= (1-\rho) \cdot \rho \cdot \frac{d}{d\rho} \sum_{i=0}^{\infty} \rho^i \\
&= (1-\rho) \cdot \rho \cdot \frac{d}{d\rho} \frac{1}{1-\rho} \\
&= (1-\rho) \cdot \rho \cdot \frac{1}{(1-\rho)^2} \\
&= \frac{\rho}{1-\rho}
\end{aligned}
$$

- Mittlere Verweilzeit (nach Little): $E(R) = \frac{E(Q)}{\lambda} = \frac{1}{\mu - \lambda}$

- Mittlere Anzahl wartender Kunden:

$$E(Q_W) = \sum_{i=1}^{\infty}(i-1)\cdot p(i) = (1-\rho)\cdot\rho\cdot\sum_{i=1}^{\infty} i\cdot\rho^i$$

$$= \rho\cdot E(Q) = \frac{\rho^2}{1-\rho}$$

- Mittlere Wartezeit: $E(R_W) = \frac{E(Q_W)}{\lambda} = \frac{\rho}{\mu\cdot(1-\rho)}$

- Varianz der Kundezahl: $\sigma^2(Q) = E(Q^2) - (E(Q))^2 = \frac{\rho}{(1-\rho)^2}$

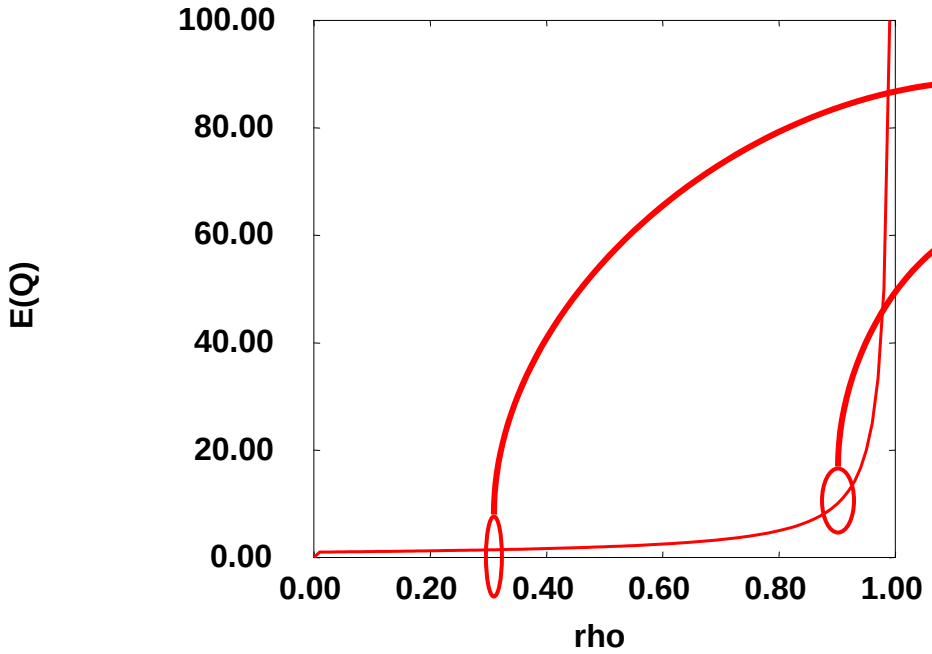
Ergebnisse für unser Beispiel:

λ	E(Q)	σ^2(Q)
0.05	0.053	0.235
0.5	1.000	1.414
0.95	19.00	19.49
0.99	99.00	99.50

Für ρ = 0.99 hatte die Simulation nach 2000 Zeiteinheiten die stationäre Phase noch lange nicht erreicht
(simulierte Population 31.7 statt 99.0)

- Da keine Annahmen über die Bedienreihenfolge gemacht wurden, gelten die Ergebnisse für alle Bedienstrategien, bei denen der Bediener arbeitet solange Kunden warten!
- Falls die Verteilung der Verweilzeit untersucht werden soll (was hier nicht geschieht), müssten wir Annahmen über die Bedienstrategie machen

Verlauf der Population:



Fast linearer Verlauf für niedrige Auslastung

Starker Anstieg für $\rho>0.8$

Struktur der Kurve ist typisch für unterschiedliche Arten von Wartesystemen, konkreter Verlauf hängt von den Parametern ab!

Ein einfaches Beispiel: Analyse eines Routers in einem Rechnernetz

Annahme: Exponentiell verteilte Zwischenankunfts- und Bedienzeiten

Messungen haben ergeben:

- Mittlere Ankunftsrate 125 Pakete pro Sekunde
- Mittlere Bedienzeit 2 msek. (= 0.002 Sekunden)

Fragestellungen:

- Wie viele Pakete sind im Mittel am Router?
- Wie groß ist die Wahscheinlichkeit, dass sich mehr als 13 Pakete am Router befinden? (Dimensionierung der Puffer)

Modellierung als M/M/1-System mit $\lambda = 125$, $\mu = 500 \Rightarrow \rho = 0.25$

Damit gilt:

$E(Q) = \rho / (1 - \rho) = 0.25 / 0.75 = 1 / 3$

$p(i) = (1 - \rho) \cdot \rho^i \Rightarrow P[i > 13] = 1 - \sum_{j=0}^{13} 0.75 \cdot 0.25^j \approx 1.49 \cdot 10^{-8}$

(Situation kommt im Mittel alle 14 Tage einmal vor!)

Weitere M/M/... Systeme

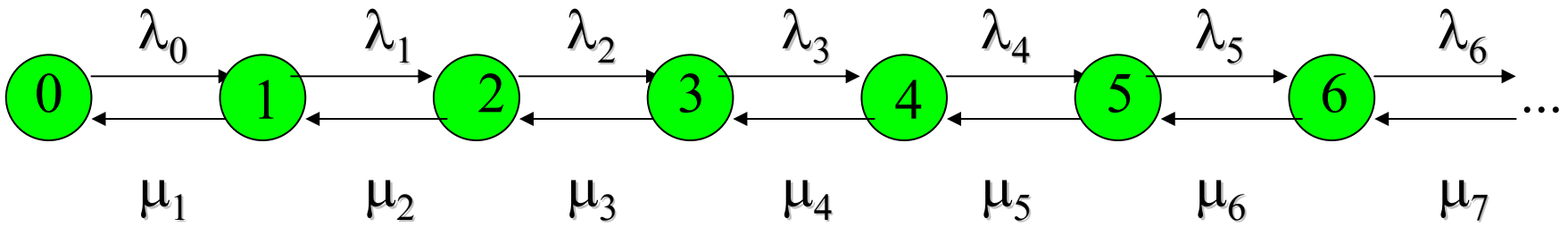
Grundsätzliches Vorgehen anwendbar auf alle Wartesysteme mit exponentiell-verteilten Zwischenankunftszeiten und Bedienzeiten.

- Zustandsraum ist (Teil-) Menge der positiven ganzen Zahlen
- Transitionen beschreiben einen Geburts-/Todesprozess

Wo liegen Unterschiede?

1. Bedienraten und Ankunftsraten können von der Population abhängen
   a. Mehrbediener-Systeme, bei denen die Bedienrate so lange linear steigt, bis alle Bediener belegt sind
   b. Systeme bei denen Kunden durch die Länge der Warteschlange abgeschreckt werden
2. Die Population oder Warteschlangenlänge kann endlich sein
3. Kombination aus 1. und 2.

Allgemeinste Form von 1. mit beliebigen Raten:



Rekursive Beziehung der Zustandswahrscheinlichkeiten:

$$p(i)\cdot\lambda_i = p(i+1)\cdot\mu_{i+1} \Rightarrow p(i) = p(0)\cdot\prod_{j=1}^{i}\frac{\lambda_{j-1}}{\mu_j}$$

und $$p(0) = \left(1+\sum_{i=1}^{\infty}\prod_{j=1}^{i}\frac{\lambda_{j-1}}{\mu_j}\right)^{-1}$$

Unendliche Summe für allgemeine Raten nicht berechenbar!

Fälle, in denen die Summe berechenbar ist:

- Endliche Anzahl lastabhängiger Raten:
  $\lambda_{i-1} / \mu_i = \lambda^* / \mu^* < 1$ für $i > n^*$ (Bsp. M/M/c)
- Fallender Quotient $\lambda_{i-1} / \mu_i < \lambda_i / \mu_{i+1} < 1$ für $i > n^*$

Berechnung M/M/c

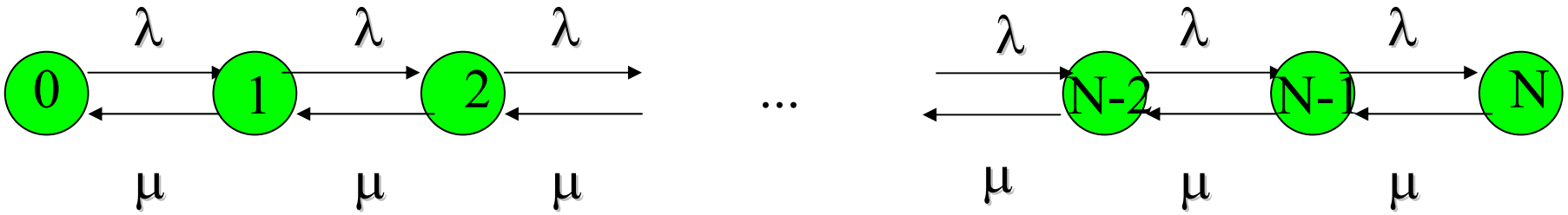
Es gilt:

$$p(i) = \begin{cases} p(0) \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{i!} & \text{falls } i \le c \\ p(0) \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{c! \cdot c^{i-c}} & \text{falls } i > c \end{cases}$$

$$\text{Mit } p(0) = \left(1 + \sum_{i=1}^{c-1} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \frac{\lambda^c}{c! \cdot \mu^c} \cdot \frac{\rho}{1-\rho}\right)^{-1} \quad \text{und } \rho = \lambda / (c \cdot \mu)$$

Modelle mit endlicher Kapazität

M/M/1/N-System



Rekursive Beziehungen bleiben, also $p(i) = \rho^i \cdot p(0)$

Falls $\rho \neq 1$ gilt: $p(0) = \frac{1-\rho}{1-\rho^{N+1}}$ und damit $p(i) = \frac{(1-\rho)\cdot\rho^i}{1-\rho^{N+1}}$

Falls $\rho = 1$ gilt: $p(i) = \frac{1}{N+1}$

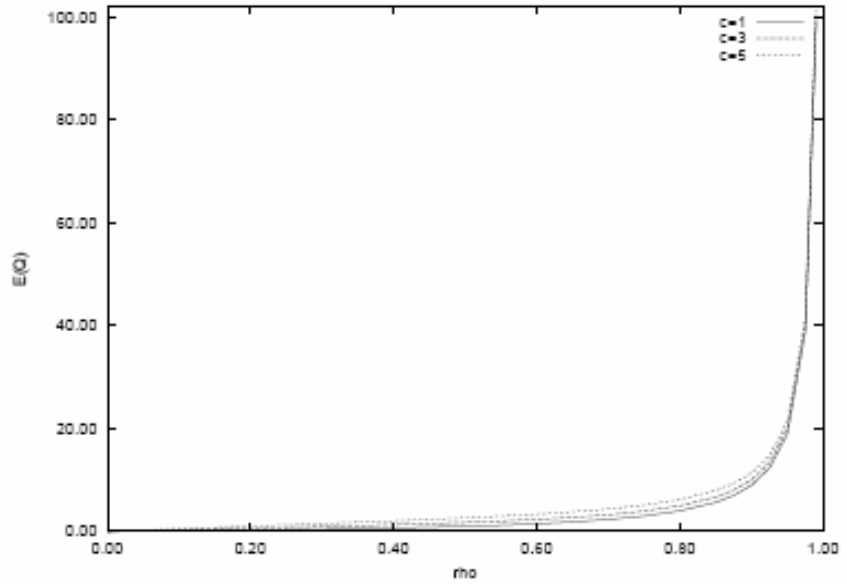
Stationäre Lösung existiert für beliebige Werte von ρ!

W. dafür, dass alle Warteplätze belegt:

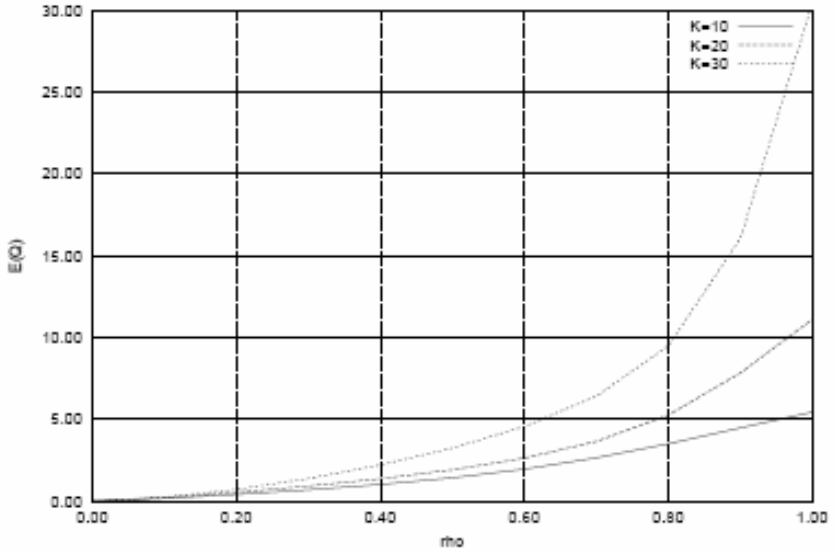
$$p(N) = \frac{(1-\rho)\cdot\rho^N}{1-\rho^{N+1}} \text{ bzw. } \frac{1}{N+1}$$

dies entspricht hier der W. einer Kundenabweisung

Population M/M/c



Population M/M/1/N



Weitere M/M/... System, für die Lösungen berechenbar sind:

- M/M/m/m nur Bediener und keine Warteplätze (Verlustsystem)
- M/M/1/./K mit K Kunden in der Umgebung jeder Kunde hat Ankunftsrate λ, damit ergibt sich Ankunftsrate $(K-i)\cdot\lambda$, wenn i Kunden am Bediener sind
- M/M/∞/./K wie vorher, nur ein Bediener pro Kunde
- M/M/m/N/K Kombination aus den vorherigen Systemen

Vorgehen bei der Lösung in allen Fällen:

1. Aufstellen der rekursiven Beziehungen zwischen den Zustandswahrscheinlichkeiten (jeweils Geburts-/Todes-Prozesse)
2. Berechnung der Zustandswahrscheinlichkeiten
3. Ermittlung der Leistungsgrößen

Dieses Vorgehen ist immer einer Simulation vorzuziehen!

Allgemeine Bedienzeiten oder Ankunftsprozesse!?

Bisherige Analyse basiert auf speziellen Eigenschaften der Exponential-Verteilung:

- Gedächtnislosigkeit
- Additivität

beides geht verloren, wenn wir andere Verteilungen verwenden

Aber Ergebnis bekannt für M/G/1

Population M/G/1: $E(Q) = \frac{\rho}{1-\rho} + \frac{\rho^2 \cdot \left(VK^2(B) - 1\right)}{2 \cdot (1-\rho)}$

Nur das erste und zweite Moment der Bedienzeit beeinflussen die Population.
Für Exp.-Vert. ist $VK^2(B)=1$.

Da der Durchsatz λ entspricht, kann E(R) über den Satz von Little berechnet werden.

Analytische/Numerische Resultate für weitere Systeme z.T. über die Analyse des unterliegenden Markov-Prozesses

(später etwas mehr dazu)

2.2.2 Offene Warteschlangennetze

Einzelne Stationen erlauben die Modellierung eines Typs von Ressourcen, die von Kunden mit (stochastisch) identischen Bedienwünschen belegt wird

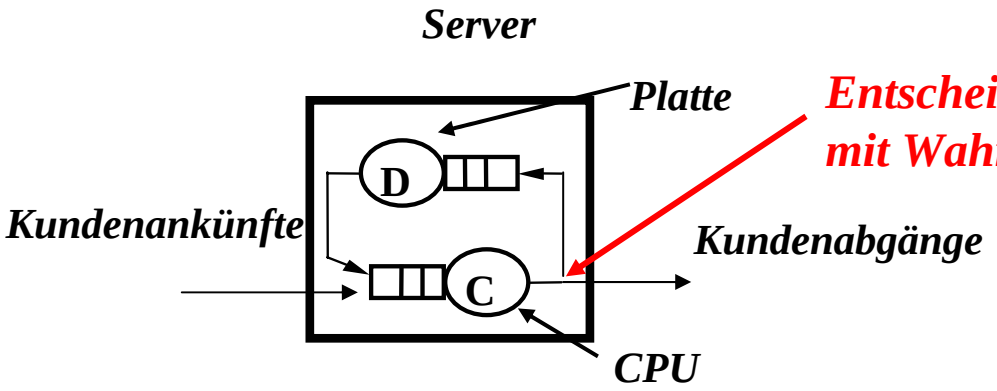
Reale Systeme zeichnen sich dadurch aus, dass

- unterschiedliche Ressourcen
  - nacheinander nach vorgegebenem Ablaufplan oder
  - mit bestimmten Übergangswahrscheinlichkeiten

  belegt werden
- Kunden mit unterschiedlichem (stochastischen) Verhalten im Netz zirkulieren, die
  - entweder permanent im System bleiben (geschlossene Kette)
  - oder das System betreten und wieder verlassen (offene Kette)

⇒ Warteschlangennetze

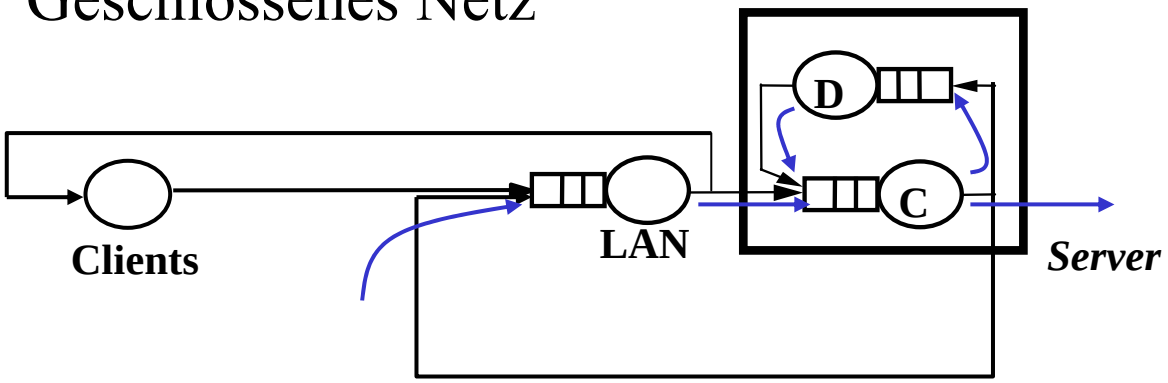
Offenes Netz



Entscheidung über den nächsten Plattenzugriff mit Wahrsch. p (Verlassen des System mit W. 1-p)

Einbeziehung der Umgebung (Beschreibung des Kundenverhaltens außerhalb des Servers)

Geschlossenes Netz



Zusätzliche Zugriffe durch Kunden von Außen

Gemischtes Netz

Wir betrachten in dieser Vorlesung nur offene Netze mit einer Kundenklasse Weitere Resultate in der Vorlesung Leistungsbewertung und Kapazitätsplanung

Beschreibung des Modelltyps

- J Stationen
- Station $j \in \{1,...,J\}$ mit unbeschränkte Kapazität und Bedienrate μ_j (Bedienzeiten sind exponentiell verteilt)
- Ankunftsrate von neuen Kunden an Station j: λ_{0j} (Zwischenankunftszeiten exponentiell-verteilt)
- Ankunftsrate an Station j: λ_j (inkl. der Kunden, die von anderen Stationen kommen)
- p_{ij} Wahrscheinlichkeit, dass ein Kunde nach Verlassen von Station i zu Station j wechselt (Routingwahrscheinlichkeiten)
- p_{i0} Wahrscheinlichkeit, dass ein Kunde nach Verlassen von Station i das Netz verlässt

Es muss gelten $\sum_{j=0}^{J} p_{ij} = 1$ für alle $i \in \{1,..., J\}$

und $\lambda_j < \mu_j$ für alle $j \in \{1, ..., J\}$ (für stationäres Gleichgewicht)

Berechnung von λ_j:

Annahme: Netz sei im stationären Gleichgewicht d.h. $\lambda_j < \mu_j$

$\Rightarrow$ Durchsatz durch Station j entspricht λ_j

Damit gilt $\lambda_j = \lambda_{0j} + \sum_{i=1}^{J} p_{ij} \cdot \lambda_i$ (lineares Gleichungssystem mit J Variablen)

In Matrixschreibweise: $\mathbf{\Lambda} = \mathbf{\Lambda} \cdot \mathbf{P} + \mathbf{\Lambda}_0 \Rightarrow \mathbf{\Lambda} = \mathbf{\Lambda}_0 \cdot (\mathbf{I} - \mathbf{P})^{-1}$

mit **P** als J×J-Matrix der Routingwahrscheinlichkeiten, ohne die Werte p_{j0}

$\mathbf{\Lambda}_0$ Vektor der Ankunftsraten λ_{0j} (bekannt)

$\mathbf{\Lambda}$ Vektor der Ankunftsraten λ_j (zu berechnen)

Lösung existiert, falls die inverse Matrix existiert

Die inverse Matrix existiert, falls die Stationen sich nicht so umordnen lassen, dass in **P** eine Untermatrix entlang der Hauptdiagonalen entsteht, in der alle Zeilensummen 1 sind!

Zustand des Netzes $(n_1,...,n_J)$ wobei n_j Population an Station j ist

$P[n_1,...,n_J]$ Wahrscheinlichkeit für Zustand $(n_1,...,n_J)$

Sei $\rho_j = \lambda_j / \mu_j$ und $P_i[n_i] = (1- \rho_i) \cdot (\rho_i)^{ni}$ (Zustandsw. M/M/1-System)

Fundamentales Resultat von Jackson (1963):

$$P[n_1,...,n_J] = \prod_{i=1}^{J} P_i[n_i]$$ (Produktformlösung!)

Resultat ist überraschend, da Zwischenankunftszeiten von Kunden an einer Station nicht zwangsläufig exponentiell verteilt sind!

Vorgehen bei der Analyse:

- Berechne die Ankunftsraten λ_j (Lsg. lineares Gleichungssystem)
- Analysiere J M/M/1-Systeme

$\Rightarrow$ Resultate für die einzelnen Stationen ($E(U_i)$, $E(X_i)$, $E(Q_i)$, $E(R_i)$)

Berechnung von Resultaten auf Netzebene
(Voraussetzung stationäre Lösung existiert):

Für den Netzdurchsatz gilt: $E(X_0) = \sum_{i=1}^{J} \lambda_{0i}$

Für die Gesamtpopulation gilt: $E(Q_0) = \sum_{i=1}^{J} E(Q_i)$

Verweilzeit im Netz nach Little: $\mathrm{E}(\mathrm{R}_0) = \mathrm{E}(\mathrm{Q}_0) / \mathrm{E}(\mathrm{X}_0)$

<u>Berechnung der gesamten Verweilzeit eines Kunden in Station j</u>:

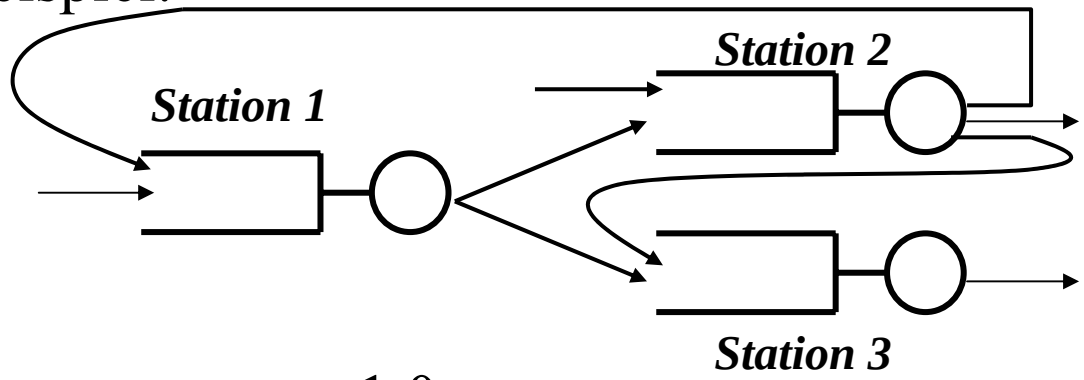
Sei $p_{0i} = \frac{\lambda_{0i}}{\sum_{k=1}^{J} \lambda_{0k}}$

$\mathbf{N} = (\mathbf{I} - \mathbf{P})^{-1}$ wobei $\mathbf{N}(i,j)$ die mittleren Anzahl von Besuchen an Station j ist, die ein Kunde, der sich gerade an Station i befindet, durchführt.

Damit gilt ergibt sich die gesamte Verweilzeit in Station j als:

$$\sum_{i=1}^{J} \mathrm{p}_{0\mathrm{i}} \cdot \left(\mathbf{N}(\mathrm{i}, \mathrm{j}) \cdot \mathrm{E}(\mathrm{R}_{\mathrm{j}}) \right)$$

Beispiel:



$\mu_1 = \mu_2 = \mu_3 = 1.0$

$\lambda_{01} = 0.4, \lambda_{02} = 0.2$

$p_{12} = p_{13} = p_{23} = 0.5, p_{21} = 0.4, p_{20} = 0.1, p_{30} = 1.0$

Gleichungssystem zur Berechnung der Ankunftsraten:

$$\left.\begin{aligned} \lambda_1 &= 0.4 + 0.4 \cdot \lambda_2 \\ \lambda_2 &= 0.2 + 0.5 \cdot \lambda_1 \\ \lambda_3 &= 0.5 \cdot \lambda_1 + 0.5 \cdot \lambda_2 \end{aligned}\right\} \quad \lambda_1 = 0.6,\ \lambda_2 = 0.5 \text{ und } \lambda_3 = 0.55$$

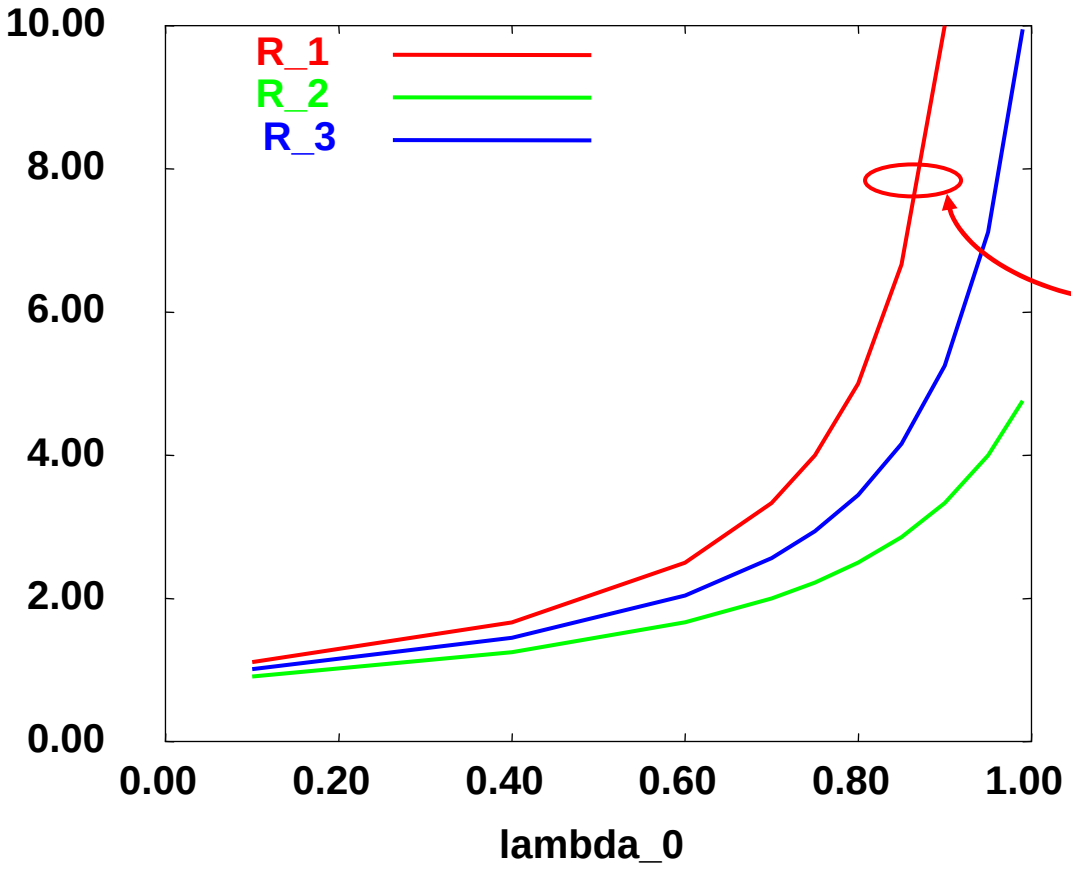
$\Rightarrow$ Analysiere jeweils ein M/M/1-System mit $\rho = 0.6, 0.5, 0.55$

$E(R_1) = 2.5, E(R_2) = 2.0, E(R_3) = 2.222$ und $E(Q_i) = E(R_i) - 1$ (i=1,2,3)

W. Netz leer: $P[n_1=0] \cdot P[n_2=0] \cdot P[n_3=0] = 0.4 \cdot 0.5 \cdot 0.45 = 0.09$

Externe Ankunftsrate $\lambda_0 = \lambda_{01} + \lambda_{02}$ und $\lambda_{01} = 2 \cdot \lambda_{02}$

Verweilzeiten an den Stationen



Verweilzeit am Falschenhals (= die am höchsten ausgelastete Station) bestimmt die Verweilzeit im Netz bei Erhöhung der Ankunftsrate!

Erweiterungen und Grenzen

- Methode ist auch anwendbar, wenn einzelne Stationen lastabhängige Bedienraten haben (und damit auch für Mehrbediener-Stationen)
  - zur Analyse auf Stationsebene dann die entsprechenden Formeln verwenden (Analyse M/M/c ...)
- Stationen mit endlicher Kapazität oder nicht exponentieller Bedienzeit sind nicht integrierbar
- Kundenklassen mit unterschiedlichem Verhalten sind mit Einschränkungen integrierbar
  (Bedienstrategie an den Stationen wird bedeutsam)
- Weitere (komplexere) Produktformresultate für geschlossene und gemischte Netze
- Zusätzlich zahlreiche auf den Produktformansätzen basierende Approximationsverfahren