Q6. Warteschlangennetze

Gliederung

1. Einfache Gesetze der Leistungsanalyse
2. Offene Netze mit einer Auftragsklasse
3. Offene Netze mit mehreren Auftragsklassen
4. Geschlossene Netze mit einer Auftragsklasse
5. Geschlossene Netze mit mehreren Auftragsklassen

Literatur (primär)

- C. G. Cassandras, S. Lafortune
  Introduction to Discrete Event Systems. Springer 2008, Kap. 8
- R. Jain
  The Art of Computer Systems Performance Analysis
  Wiley 1991, Kap. 30-34
- D. A. Menasce, V. A. F. Almeida, L. W. Dowdy. Performance by Design. Prentice Hall 2004. Kap. 11-13.
- N. J. Gunther. Analyzing Computer System Performance with Perl::PDQ. Springer 2005. Kap. 2,3.
- W. J. Stewart
  Probability, Markov Chains, Queues, and Simulation.
  Princeton University Press 2009, Part III

Software: http://jmt.sourceforge.net/Download.html

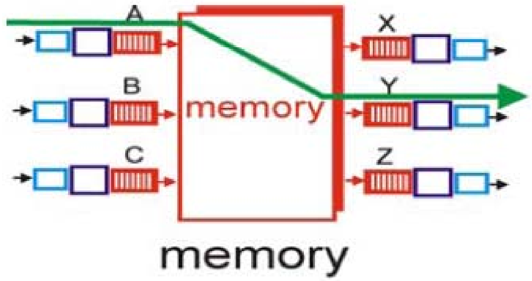
Ziele:

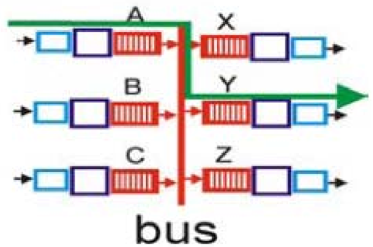
- Einige grundlegende Gesetzte der Leistungsanalyse kennen und anwenden lernen
- Elementare Warteschlangensysteme klassifizieren können
- Systeme mit Warteschlangennetzen modellieren können
- Grundlagen der Mittelwertanalyse und daraus resultierender Algorithmen kennen lernen

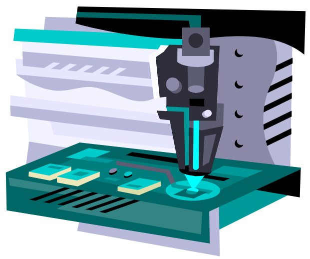
6.1 Einfache Gesetze der Leistungsanalyse

Viele Systeme lassen sich durch Aufträge/Kunden/Werkstücke/Jobs/… beschreiben, die Ressourcen/Maschinen/Schalter/… belegen

- Jeder Kunde hat einen Bedienwunsch
- Ressourcen sind nur in beschränktem Umfang vorhanden, d.h. Kunden müssen u.U. auf Zuteilung warten



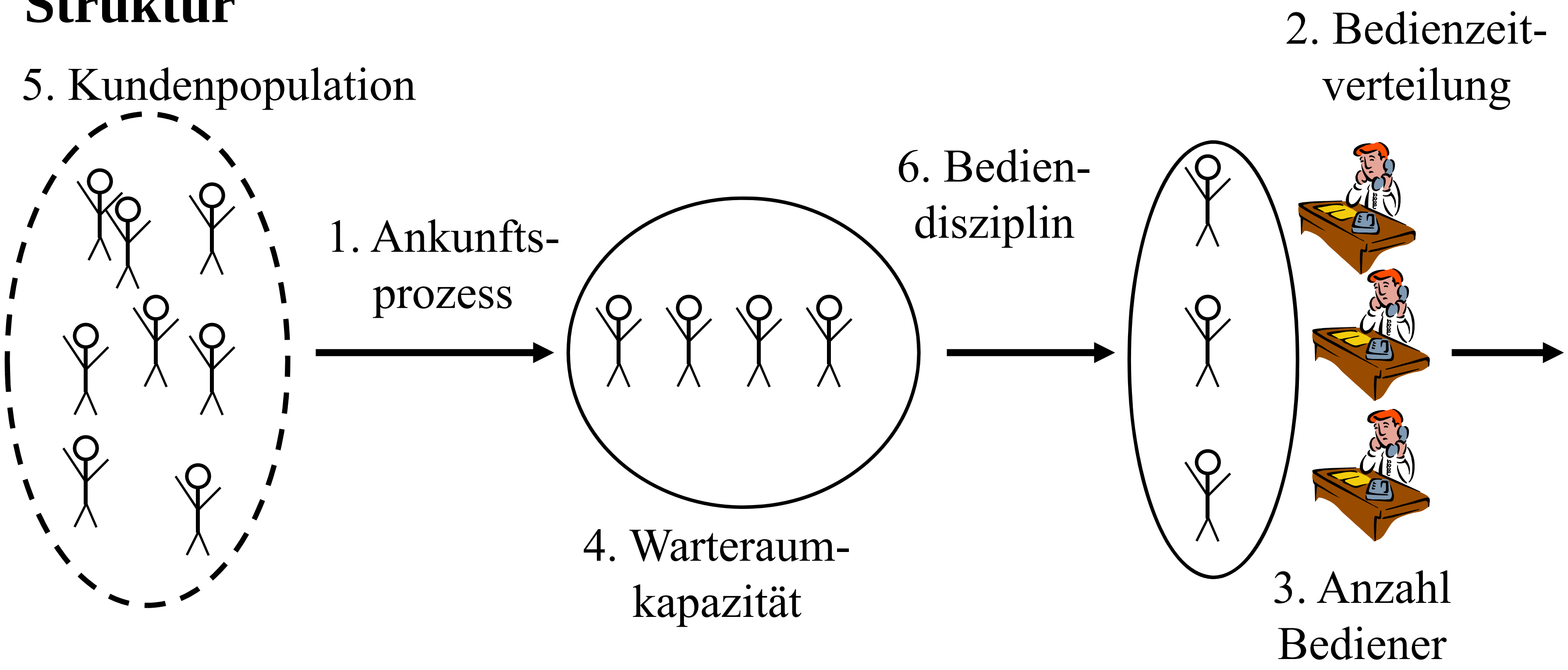




Basissystem bekannt unter mehreren Bezeichnungen Warteschlangen, Bediensysteme, Wartesysteme, Stationen, ... (+ diverse engl. Bez.)

Abstrakte Darstellung

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 < $\propto$,
      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 $\propto$)
- K Gesamtpopulation (Voreinstellung $\propto$)
- SD Bedienstrategie (Voreinstellung FCFS)

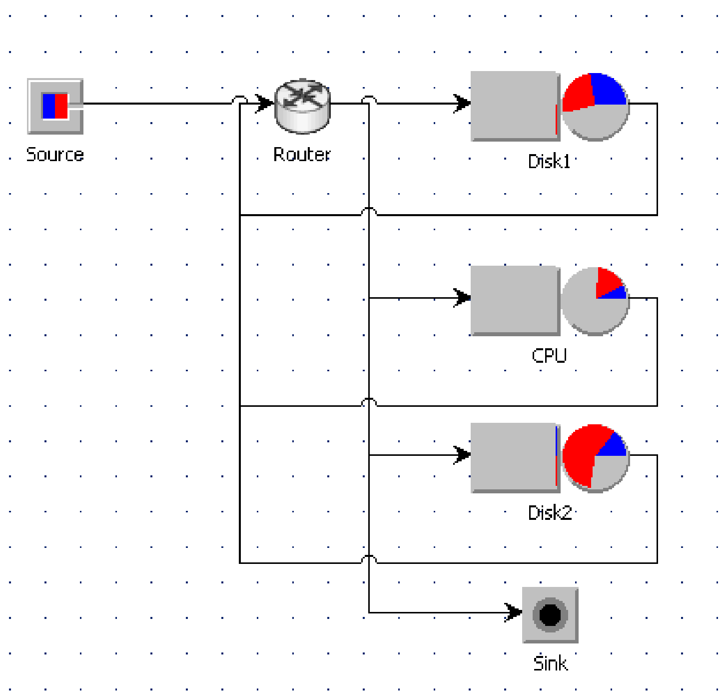
Beispiele:

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

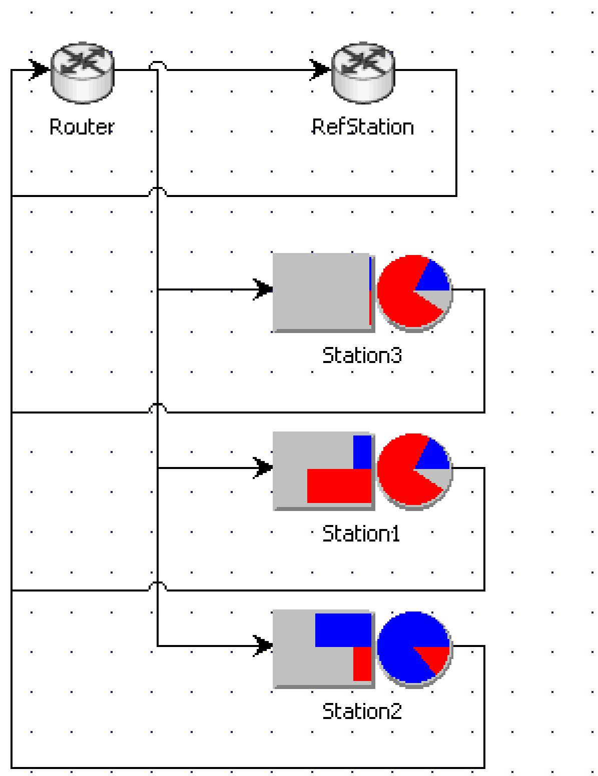
M/GI/2/5/10/RANDOM exponentiell-verteilte Zwischenankunftszeiten, unabhängig allgemein-verteilte Bedienzeiten, 2 Bediener, Warteraum mit Kapazität 5, Gesamtpopulation 10 und zufällige Auswahl der Aufträge

Warteschlangennetze bestehen aus mehreren Ressourcen, die von Aufträgen u.U. mehrfach besucht werden

Offenes System



Geschlossenes System



Bilder aus dem Tool JMT!

Analysemöglichkeiten für Warteschlangenetze:

- Ereignisdiskrete Simulation
- Analyse des zeitkontinuierlichen Markov-Prozesses (falls Zustandsraum endlich und nicht zu groß oder unendlich mit rekursiver Struktur)
- „Analytisch“ mittels Mittelwertanalyse
  - Exakt für so genannte Produktformnetze
  - Approximation sonst

Basis der Analyse: Operationale Zusammenhänge

- Verallgemeinerung von Beobachtungen endlicher Intervalle auf „typisches Verhalten“
- Gültigkeit des „Ankunftstheorems“

Prinzip der Analyse von Warteschlangensystemen:

- Stationäre Analyse auf Basis von Mittelwerten (über einen sehr langen Zeitraum)

Notationen:

V_i mittlere Anzahl Besuche an Ressource i

S_i mittlere Bedienzeit an Ressource i (Bedienrate $\mu_i = (S_i)^{-1}$

W_i mittlere Wartezeit an Ressource i

R_i mittlere Antwortzeit an Ressource i

λ_i Ankunftsrate an Ressource i

X_i mittlerer Durchsatz von Ressource i

X_0 mittlerer Durchsatz des Gesamtsystems

N_i mittlere Anzahl Aufträge an Ressource i

$N_{i,w}$ mittlere Anzahl Aufträge im Warteraum vor Ressource i

U_i Auslastung von Ressource i

Gesetze der operationalen Analyse

Es gehen keine Aufträge verloren $\Rightarrow X_i = \lambda_i$ (Durchsatz = Ankunftsrate)

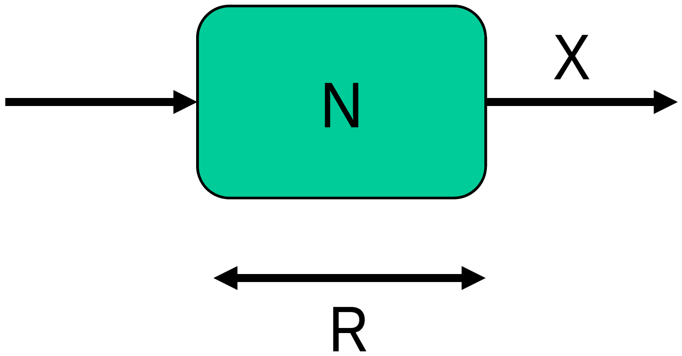
falls $\lambda_i \geq \mu_i$ System überlastet, keine stationäre Analyse möglich

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.



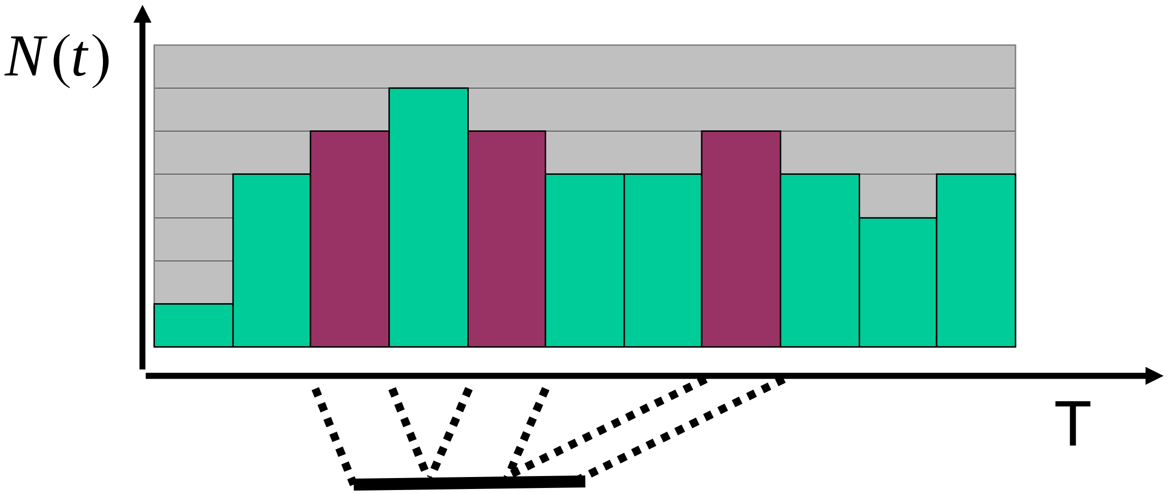
Beobachtete Größen:

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

Es gilt $N = X \cdot R$

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

- Systemzustand im Beobachtungsintervall [0,T]



- f_i prozentualer Anteil am Intervall [0,T] zu dem sich i Kunden im System befinden
- r_i Zeitdauer, zu der sich i Kunden im System befinden, damit gilt $r_i = f_i \cdot T$
- Weiterhin gilt $N = \sum_i i \cdot f_i = \sum_i i \cdot r_i / T$
- C_0 Gesamtzahl Kunden, die das System im Beobachtungsintervall verlassen haben

$$N = \frac{C_0}{T} \frac{\sum_i i \cdot r_i}{C_0} = X \cdot R$$

Gesetzt von Little gilt für allgemeine Systeme, in denen keine Aufträge zerstört oder erzeugt werden!

$\Rightarrow$ Hilfreich bei Messungen und Analysen!

Beispiele:

- LAN Router hat Latenzzeit von 200 μs/ Paket bei einer Übertragungsrate von 30000 Paketen/s.
  Mittlere Paketzahl im Router ?
  N = 30000 · 0.0002 = 6 Pakete
- Berechnung der Verweilzeit im M/M/1 System
  X = λ und N = ρ / (1 – ρ) wobei ρ = λ / μ

$$\Rightarrow \quad N = \frac{\rho}{1-\rho}\frac{1}{\lambda} = \frac{1}{1-\rho}\frac{1}{\mu} = \frac{1}{\mu-\lambda}$$

Weitere operationale Gesetze:

(für eine Station mit Index i und das gesamte Netz mit Index 0)

- Flussrelation $X_i = V_i \cdot X_0$ wobei $X_0 = \lambda$ solange das Netz nicht überlastet
- Bedienbedarf an einer Ressource $D_i = V_i \cdot S_i = (X_i / X_0) \cdot (U_i / X_i) = U_i / X_0$
- Populationsaufteilung $N_i = N_{i,w} + U_i$
- Varianten des Gesetzes von Little
  - Für den Warteraum $N_{i,w} = X_i \cdot W_i$
  - Für den Bediener $U_i = X_i \cdot S_i = \lambda_i / \mu_i = \rho_i$ (bei einem Bediener!!)

Q6.2 Offene Netze mit einer Auftragsklasse

Verschiedene Spezifikationsansätze existieren

Wir betrachten die direkte Spezifikation von

- Externer Ankunftsrate λ
- I Stationen mit
  - Besuchshäufigkeiten V_i
  - Bedienbedarfen S_i

und nehmen an, dass alle Ressourcen einen Bediener haben und alle Aufträge sich „statistisch" identisch verhalten (eine Auftragsklasse)
(es gibt andere Spezifikationsansätze)

Topologie des Netzes spielt keine Rolle, sondern nur die Besuchsanzahlen!

Zentrales Theorem für offene Netze:

Ankunftstheorem: Ein Auftrag, der an einer Station i ankommt, trifft dort im Mittel N_i Aufträge an.

Nach den operationalen Gesetzen gilt: $X_i = V_i \cdot \lambda$ und $U_i = X_i \cdot S_i$

Ferner gilt: $R_i = S_i + N_i \cdot S_i = S_i + X_i \cdot R_i \cdot S_i$ (wg. $N_i = X_i \cdot R_i$)

Implizite Annahme mittlere Verweilzeit entspricht eigener Bedienzeit + gesamter Bedienzeit der N_i anwesenden Aufträge

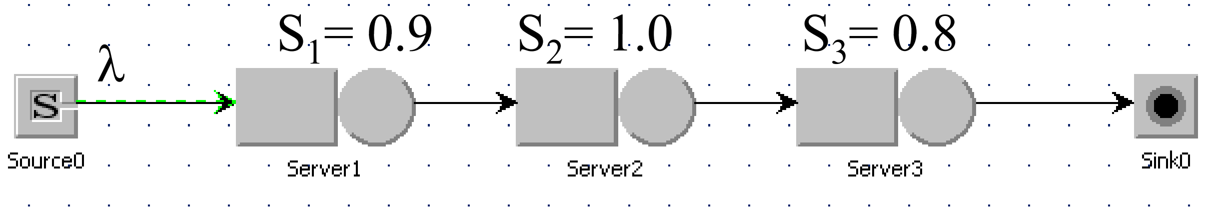
Da $U_i = X_i \cdot S_i$ gilt auch $R_i = S_i / (1 - U_i)$

damit sind R_i und N_i aus den Netzparametern berechenbar

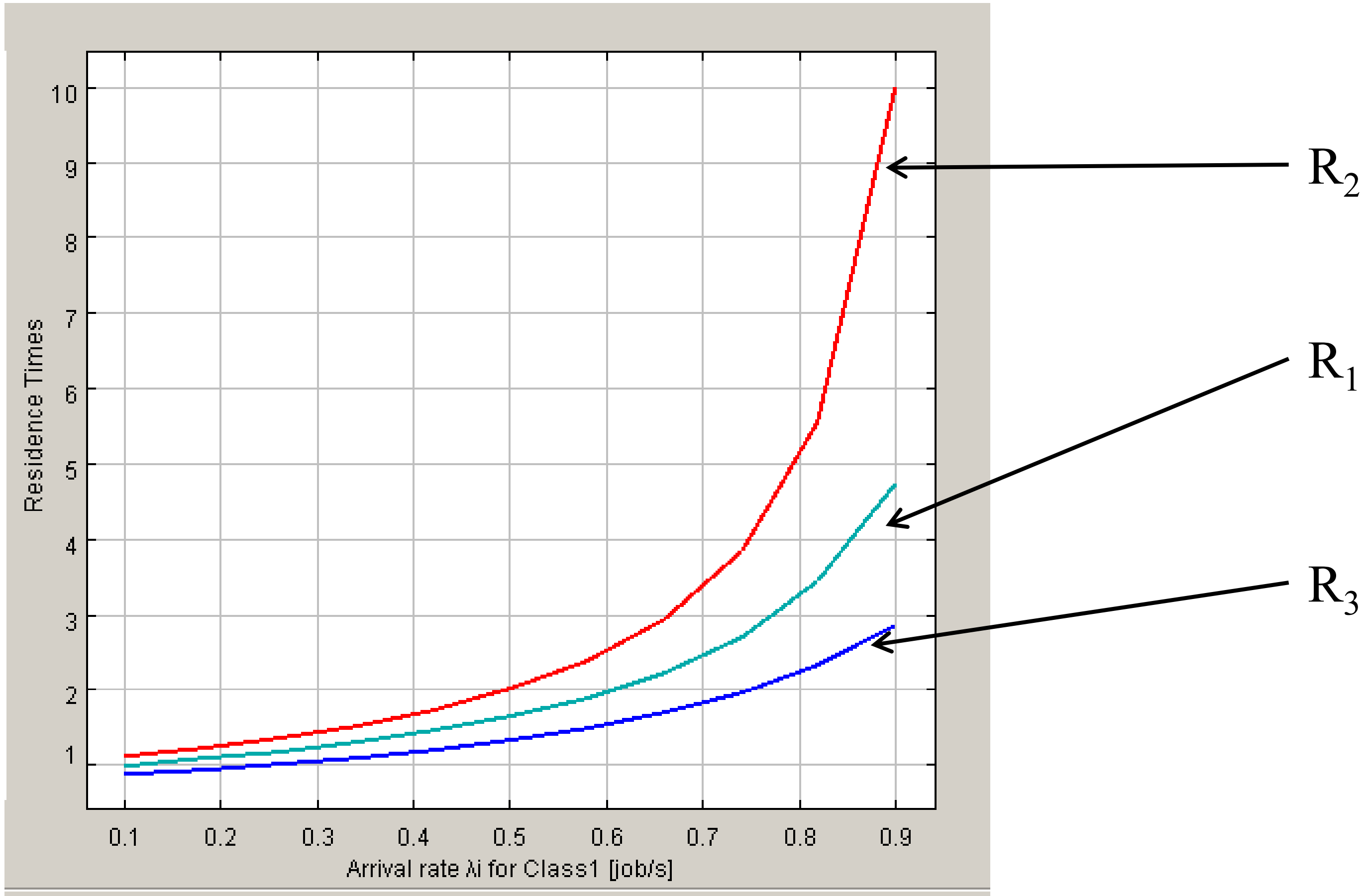
Sei $D_i = V_i \cdot S_i$ und $R'_i = V_i \cdot R_i$ (Gesamtbedienzeit und Gesamtverweilzeit)

Erlaubte Ankunftsraten: $\lambda < (\max_{i=1,\ldots,I} D_i)^{-1}$

Beispiel



S_1= 0.9
S_2= 1.0
S_3= 0.8
λ
S
Source0
Server1
Server2
Server3
Sink0



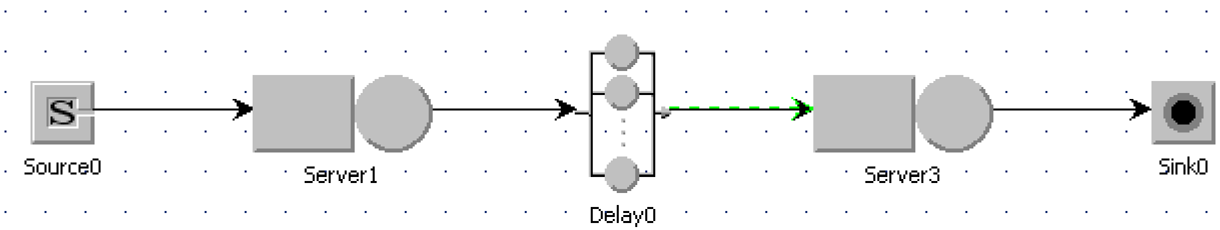
Residence Times
10
9
8
7
6
5
4
3
2
1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Arrival rate λi for Class1 [job/s]
R_2
R_1
R_3

Netzweite Leistungsgrößen

Gesamtverweilzeit $R_0 = \sum_{i=1}^{I} R'_i$ Gesamtpopulation $N_0 = \sum_{i=1}^{I} N_i$

Weitere Stationstypen sind integrierbar
(sofern der Zusammenhang zwischen R_i und N_i dargestellt werden kann)

Beispiel: Verzögerungsstation (d.h. jeder Auftrag bekommt seinen Bediener)



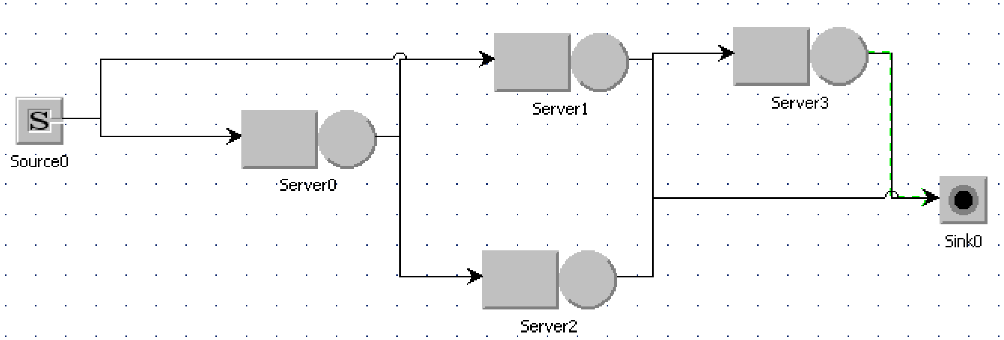
$R_i = S_i$
unabhängig von N_i

Bisherige einfachste Variante: Sequenz von Stationen (Tandemnetz)

- Jeder Kunde besucht jede Station genau einmal

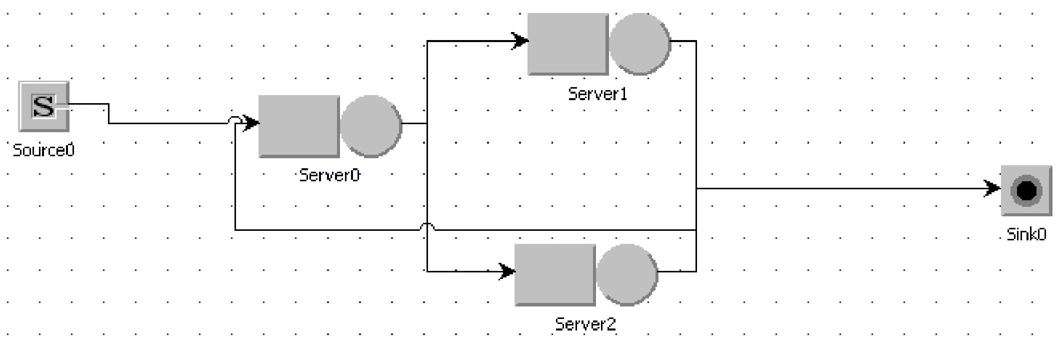
Feedforward-Netz

- Jeder Kunde besucht jede Station höchstens einmal



Allgemeines Netz

- nur Besuchshäufigkeiten zählen bei der Berechnung (siehe Folie 17)



Q6.3 Offene Netze mit mehreren Auftragsklassen

<u>Annahmen:</u>

- J unterschiedliche Verhaltensmuster (Klassen)
- Jede Klasse k ist charakterisiert durch
  - Ankunftsrate λ_k
  - Besuchshäufigkeiten V_{ik} von Station i
  - Bedienzeiten S_{ik} an Station i (Sofern Station i besucht wird, d.h. $V_{ik} > 0$) auch hier wieder $D_{ik} = V_{ik} \cdot S_{ik}$

Leistungsgrößen

- Für Klassen an Stationen (N_{ik}, R_{ik}, R'_{ik}, U_{ik})
- Für Stationen (N_i, U_i)
- Für Klassen (N_k, R_k)
- Für das gesamte Netz (N_0)

Ergebnisse lassen sich einfach auf mehrere Klassen übertragen!

$X_{ik} = V_{ik} \cdot \lambda_k$ und $U_{ik} = X_{ik} \cdot S_{ik}$ nach den operationalen Gesetzen

Ankunftstheorem:
Ein ankommender Auftrag trifft bei seiner Ankunft an einer Station die mittlere Population in allen Auftragsklassen an!

Annahme zur Berechnung der Verweilzeit:
Wenn ein Auftrag insgesamt N_i Aufträge in Station i antrifft, verzögert sich seine Verweilzeit um den Faktor $N_i \cdot S_{ik}$:

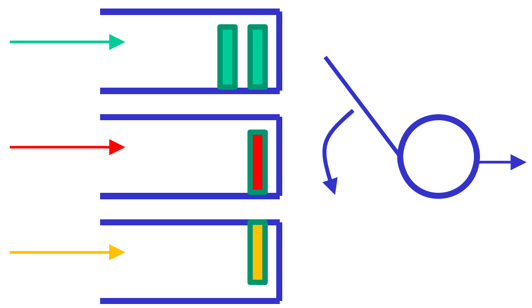
$$R_{ik} = S_{ik} + \Sigma_{l=1,\ldots,K} N_{il} \cdot S_{ik} \Rightarrow R_{ik} = S_{ik} / (1 - \Sigma_{l=1,\ldots K} U_{il}) \Rightarrow$$

$$N_{ik} = X_{ik} \cdot R_{ik} = U_{ik} / \ (1 - \Sigma_{l=1,\ldots K} U_{il})$$

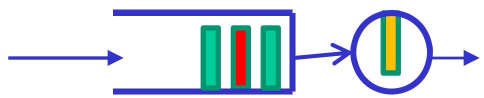
$$R'_{ik} = V_{ik} \cdot R_{ik}$$

Netz ist stationär, falls $\max_{i=1,\ldots I} \Sigma_{k=1,\ldots,K} U_{ik} < 1$

Verweilzeitberechnung



Gleichzeitige Bedienung aller Kunden mit 1/N der Kapazität
(im Beispiel ¼ Kapazität pro Kunde)
Processor Sharing



Kunden werden nach Ankunftsreihenfolge bedient (FCFS)

$$R_{ik} = S_{ik} + \Sigma_{l=1,\ldots,K} N_{il} \cdot S_{il}$$

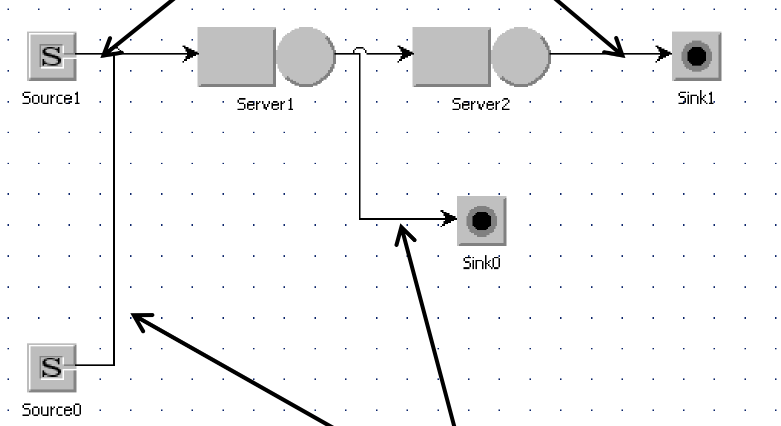
Nur Approximation!

Weitere Strategien wie Prioritäten oder Generalized PS sind auch abbildbar und approximativ analysierbar!

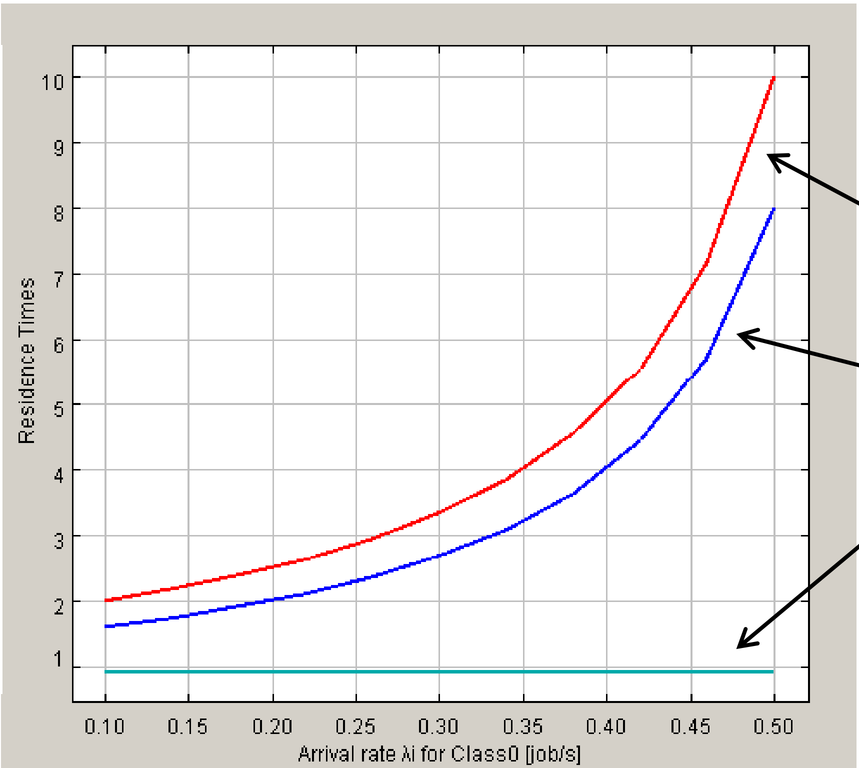
Weitere Größen leicht ableitbar:

- $U_i = \Sigma_{k=1,..,K}\, U_{ik}$ und $N_i = \Sigma_{k=1,..,K}\, N_{ik}$,
- $N_k = \Sigma_{i=1,..,I}\, N_{ik}$ und $R_k = \Sigma_{i=1,..,I}\, R_{ik}$
- $N_0 = \Sigma_{i=1,..,I}\, \Sigma_{k=1,..,K}\, N_{ik}$

Klasse 2 (λ_2=0.5)



Klasse 1 (λ_1=0.1-0.5)



R_{11}

R_{12}

R_{22}

S_{11}=1.0 S_{12}=0.8

S_{22}=1.25

Q6.4 Geschlossene Netze mit einer Auftragsklasse

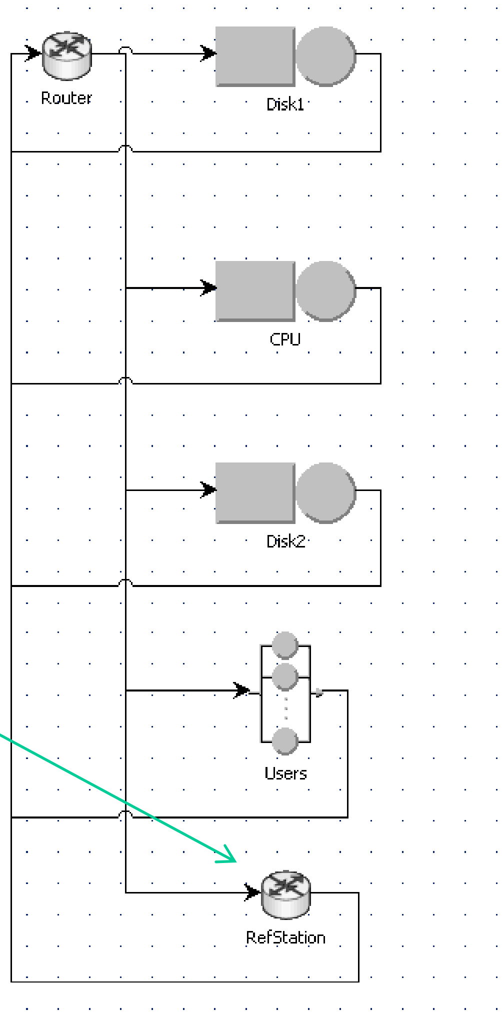
Feste Kundenpopulation im Netz

⇒ Anzahl Besuche an einer Station ∞ oder 0

⇒ Relative Besuchszahl V_i bzgl. einer Referenzstation

Wir betrachten erst einmal nur eine Auftragsklasse mit Kundenzahl N und Bedienbedarf S_i an Station i

Ziel: Berechnung X_i, N_i, R_i, U_i, sowie X_0, R_0



Leistungsgrößen in geschlossenen Netzen hängen von der Population ab

Wir definieren deshalb:

$N_i(n)$, $R_i(n)$, $U_i(n)$ mittlere Population/Antwortzeit/Auslastung von Station i, bei Population n

$X_i(n)$, $X_0(n)$ mittlerer Durchsatz von Station i/des Netzes bei Population n

$R_0(n)$ Antwortzeit des Systems bei Population n

Sei ferner $N_{ai}(n)$, die mittlere Population, die ein Auftrag antrifft, wenn er an Station i ankommt

Grundlage der Analyse: Ankunftstheorem (Reiser/Lavenberg 1980)

Ankunftstheorem (für geschlossene Netze):
Ein ankommender Auftrag in einem Netz mit n Aufträgen trifft bei seiner Ankunft an Station i die mittlere Population an dieser Station mit einem Auftrag weniger im Netz an!
D.h. $N_{ai}(n) = N_i(n-1)$

Wenn die Population bei Ankunft bekannt ist, kann daraus die Antwortzeit berechnet werden

- $R_i(n) = S_i + N_{ai}(n) \cdot S_i = S_i + N_i(n-1) \cdot S_i$ für Stationen mit einem Bediener
- $R_i(n) = S_i$ für Verzögerungsstationen

Die Gesamtverweilzeit an Station i und im Netz entspricht dann

- $R'_i(n) = V_i \cdot R_i(n)$ und $R_0 = \Sigma_{i=1,\ldots,I} R'_i$

Daraus kann man den Durchsatz des Netzes und jeder Station berechnen

- $X_0(n) = n / R_0(n)$ und $X_i(n) = V_i \cdot X_0(n)$

Mit Hilfe des Satzes von Little folgt dann

- $N_i(n) = X_i(n) \cdot R_i(n) = X_0(n) \cdot R'_i(n)$

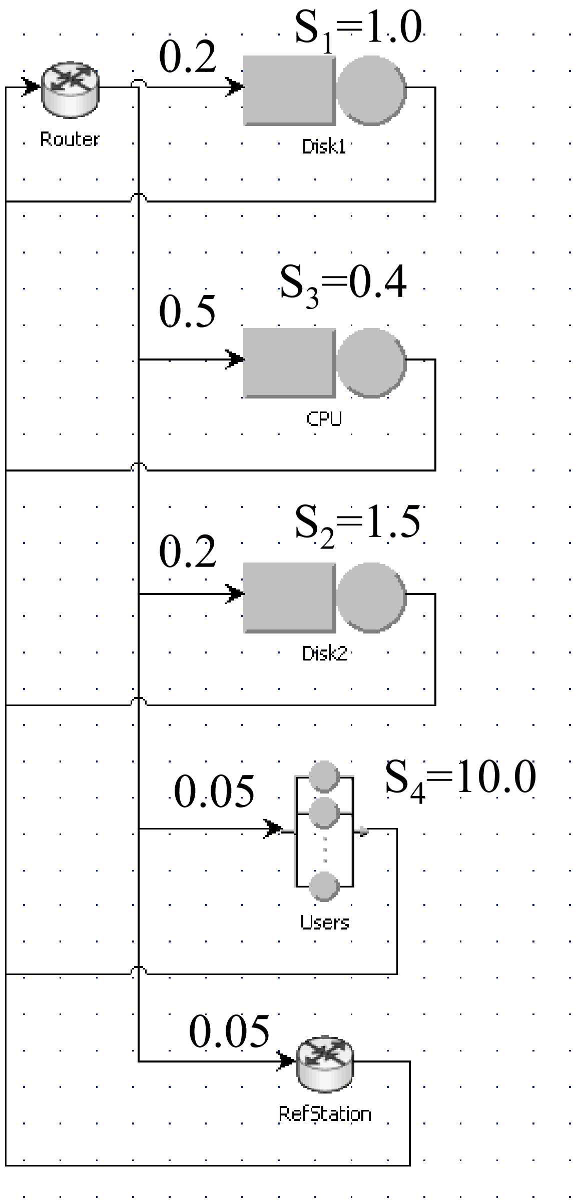
damit ist dann

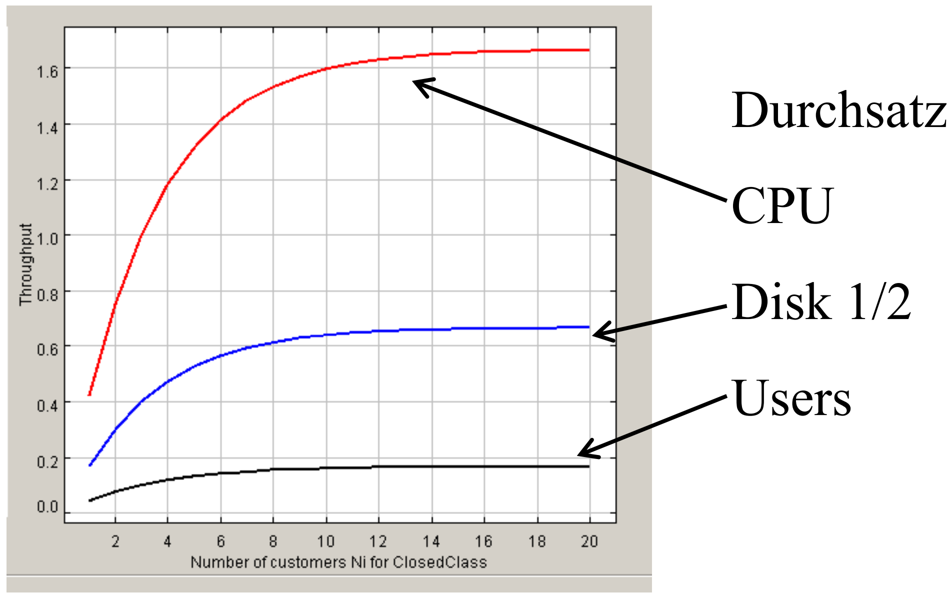
- $N_{ai}(n+1) = N_i(n)$

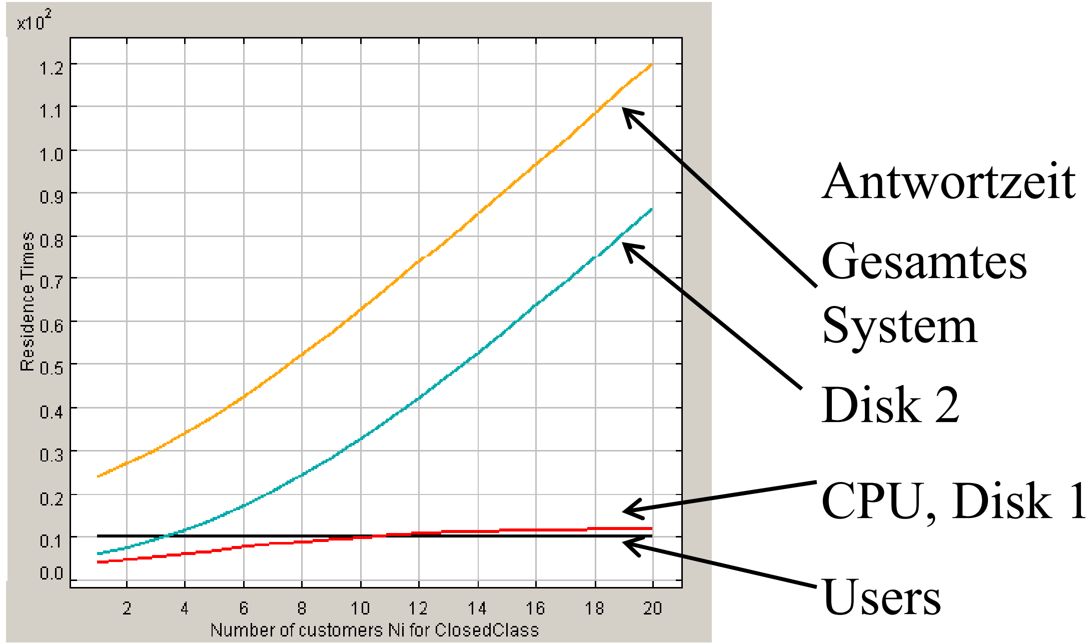
n=n+1

bekannt

$\Rightarrow$ Algorithmus der von n=0 (mit $N_i(0)=0$) bis n=N läuft


0.2
S_1=1.0
Router
Disk1
0.5
S_3=0.4
CPU
0.2
S_2=1.5
Disk2
0.05
S_4=10.0
Users
0.05
RefStation


Throughput
Number of customers Ni for ClosedClass
Durchsatz
CPU
Disk 1/2
Users


x10²
Residence Times
Number of customers Ni for ClosedClass
Antwortzeit
Gesamtes
System
Disk 2
CPU, Disk 1
Users

Q6.5 Geschlossene Netze mit mehreren Auftragsklassen

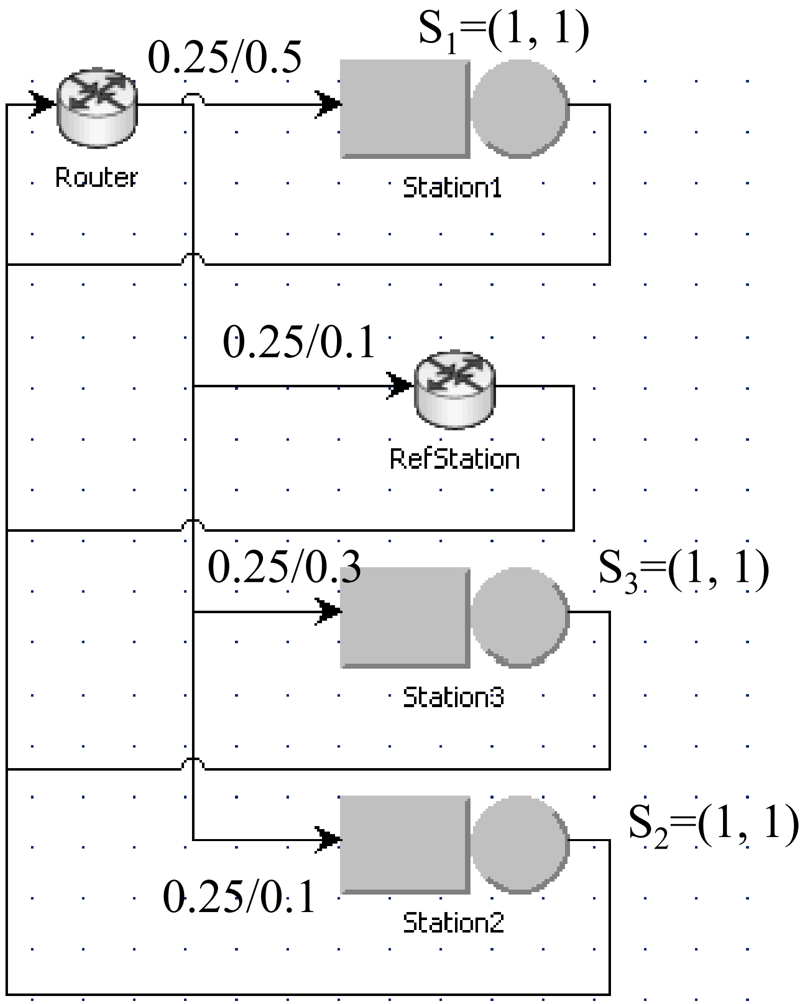
Wie im Fall offener Netze K unterschiedliche Verhaltensmuster (Klassen) mit spezifischen

- Besuchshäufigkeiten V_{ik}
- Bedienzeiten S_{ik}

Leistungsgrößen nun in Abhängigkeit vom Populationsvektor
$\mathbf{n} = (n_1, \ldots, n_K)$

Leistungsgrößen

- $R_{ik}(\mathbf{n})$, $X_{ik}(\mathbf{n})$, $U_{ik}(\mathbf{n})$, $N_{ik}(\mathbf{n})$
- $X_{0k}(\mathbf{n})$, $R_{0k}(\mathbf{n})$



Ankunftstheorem (für geschlossene Netze mit mehreren Klassen):
Ein ankommender Auftrag der Klasse k in einem Netz mit **n** Aufträgen trifft bei seiner Ankunft an Station i die mittlere Population an dieser Station mit einem Auftrag der Klasse k weniger im Netz an!
D.h. $N_{aik}(\mathbf{n}) = N_{ik}(\mathbf{n-e}_k)$
(mit $\mathbf{e}_k$ Vektor der Länge K mit 1 in Position k und 0 sonst)

Berechnung der Leistungsgrößen:

- $R_{ik}(\mathbf{n}) = S_{ik} + \Sigma_{l=1,\ldots,K} N_{ail}(\mathbf{n}) \cdot S_{ik} = S_{ik} + \Sigma_{l=1,\ldots,K} N_{il}(\mathbf{n-e}_k) \cdot S_{ik}$
  für Stationen mit einem Bediener
  $R_{ik}(\mathbf{n}) = S_{ik}$ für Verzögerungsstationen
- $R'_{ik}(\mathbf{n}) = V_{ik} \cdot R_i(\mathbf{n})$ und $R_{0k} = \Sigma_{i=1,\ldots,I} R'_{ik}$
- $X_{0k}(\mathbf{n}) = n_k / R_{0k}(\mathbf{n})$ und $X_{ik}(\mathbf{n}) = V_{ik} \cdot X_{0k}(\mathbf{n})$
- $N_{ik}(n) = X_{ik}(\mathbf{n}) \cdot R_{ik}(\mathbf{n}) = X_{0k}(\mathbf{n}) \cdot R'_{ik}(\mathbf{n})$ und $N_i(\mathbf{n}) = \Sigma_{k=1,\ldots,K} N_{ik}(\mathbf{n})$
- $N_{aik}(\mathbf{n+e}_k) = N_{ik}(\mathbf{n})$

Algorithmus zur Berechnung:

$N_{ik}(\mathbf{0}) = N_i(\mathbf{0}) = \mathbf{0}$;

For ($n_1 = 0$ to N_1) {

….

For ($n_K = 0$ to N_K) {

For ($k = 1$ to K) {

if ($n_k > 0$) {

For ($i = 1$ to I) {

$R_{ik}(\mathbf{n}) = S_{ik} \cdot (1 + N_i(\mathbf{n}\text{-}\mathbf{e}_k))$; }

$X_{0k}(\mathbf{n}) = n_k / (\Sigma_{i=1,\dots I} V_{ik} \cdot R_{ik}(\mathbf{n}))$;

$N_{ik}(n) = X_{0k}(\mathbf{n}) \cdot V_{ik} \cdot R_{ik}(\mathbf{n})$ }

$N_i(\mathbf{n}) = \Sigma_{k=1,\dots K} N_{ik}(\mathbf{n})$; } … }

Aufwand:

Zeit O(…)

Speicher O(…)

Für große Populationen und Klassenzahlen kann der Aufwand zu hoch sein

$\Rightarrow$Approximationen zur effizienteren Berechnung

Es existieren zahlreiche Varianten

Ansatz von Schweitzer (erster Approximationsansatz in diesem Kontext)

Anzahl der Aufträge in einer Station steigt proportional zur Auftragszahl

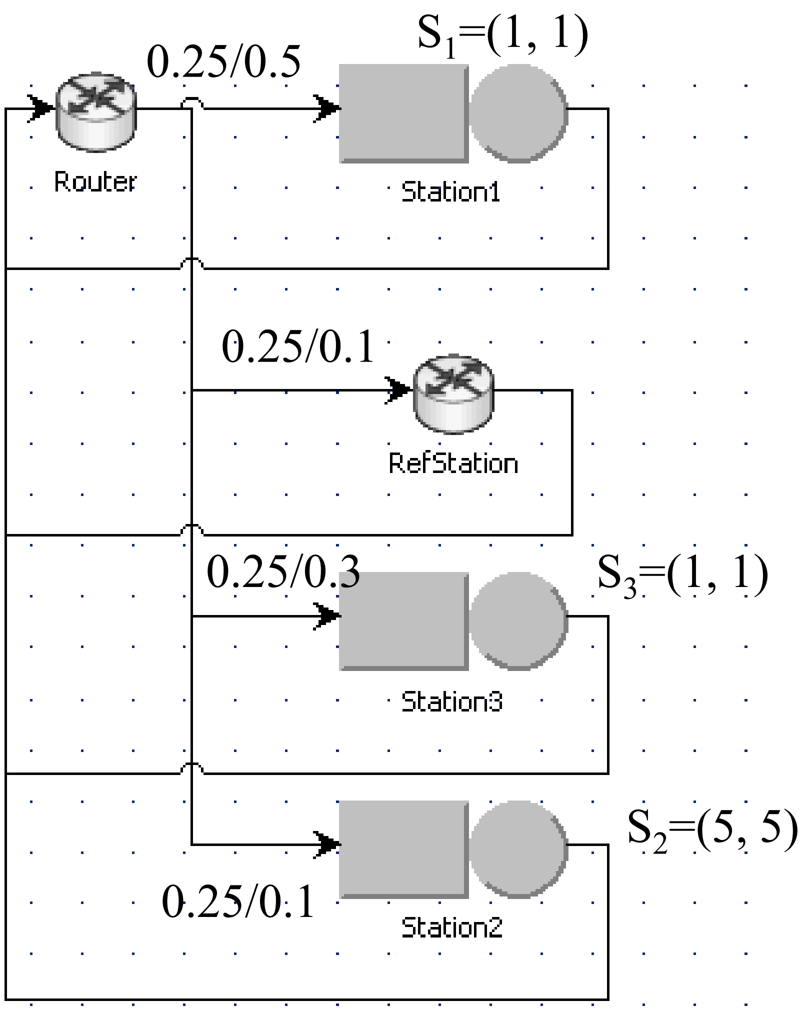
$N_{ik}(\mathbf{n}\text{-}\mathbf{e}_k) / N_{ik}(\mathbf{n}) = (N_k - 1) / N_k$

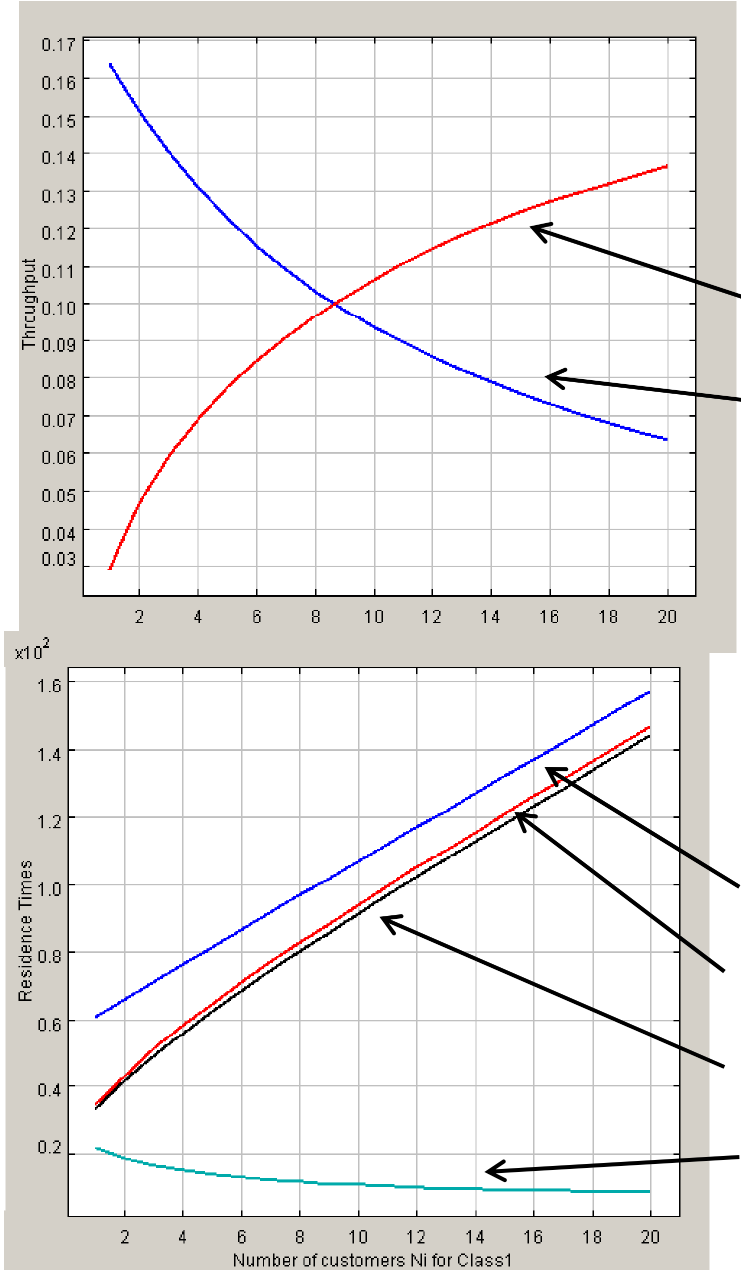
Fixpunktansatz:

1. Initialisiere $N_{ik}(\mathbf{N})$;
2. Berechne $N_{ik}(\mathbf{N}\text{-}\mathbf{e}_k) = N_{ik}(\mathbf{N}) \cdot (N_k - 1) / N_k$;
3. Bestimme $R_{ik}(\mathbf{N})$ und $X_{0k}(\mathbf{N})$ wie vorher ;
4. Berechne $N'_{ik}(\mathbf{N}) = X_{0k}(\mathbf{N}) \cdot V_{ik} \cdot R_{ik}(\mathbf{N})$;
5. Falls $\max_{i,k} | N'_{ik}(\mathbf{N}) - N_{ik}(\mathbf{N})| < \varepsilon$
   beende Berechnung
6. $N_{ik}(\mathbf{N}) = N'_{ik}(\mathbf{N})$ und fahre bei 2. fort ;

Konvergenz nicht bewiesen (nichtlineares Fixpunktproblem), i.d.R. aber Konvergenz nach wenigen Iterationen mit kleinen Approximationsfehlern!

Population Klasse 1: 1-20, Klasse 2: 10





Durchsatz

Klasse 1

Klasse 2

Antwortzeit

Klasse 2

Klasse 1

Klasse 2/St. 2

Klasse 2/ St. 1

Idee der Mittelwertanalyse:

„Ankommender Auftrag sieht den stationären Zustand“

ermöglicht die Realisierung effizienter Analysealgorithmen für viele Modelle

- Resultate sind für spezielle Modellklassen exakt
- i.a. aber nur Approximationen (oft mit nur geringen Fehlern)
- insgesamt eine weites Feld mit vielen Anwendungen
- oft eine Alternative zur Simulation