2. Mathematische Leistungsmodelle

Problemfeld

Erinnern wir uns: Wir betrachten eine gewisse Menge von Rechensystemen (MASCHINEN) sowie eine gewisse Menge von Rechenlasten (LASTEN), die von diesen Rechensystemen bearbeitet werden können. Wir einigen uns ferner darauf, wie wir die "Güte der Bearbeitung einer Last durch ein System" bewerten wollen, und legen gewisse (dieser Absicht angepaßte) Leistungsmaße samt einer zugeordneten Menge von Leistungswerten (L-WERTE) fest. Unser Ziel ist es, den Zusammenhang zwischen den verschiedenen (Maschine, Last)-Paaren und den jeweils zugehörigen Leistungswerten zu ermitteln. Wir hatten diesen Zusammenhang gekennzeichnet durch die "Leistungsfunktion"

L: MASCHINEN x LASTEN —> L-WERTE

deren genaue Kenntnis uns die Beantwortung aller im Leistungsbereich relevanten Fragen ermöglichen würde.

Objekt-experimente

Wir haben erwähnt, daß wir mittels Objektexperimenten die Funktion L punktweise abtasten könnten. Zu diesem Zweck müßten wir interessierende Rechensysteme (bzw. Rechensystem-Versionen) jeweils einzeln de facto installieren/implementieren, sie mit interessierenden Lasten (bzw. Last-Versionen) jeweils einzeln de facto beschicken, den Vorgang der Bearbeitung einer Last durch ein System mittels geeigneter Meßinstrumente jeweils einzeln beobachten und schließlich aus den Beobachtungen den gemäß gewählter Leistungsmaße jedem (System, Last)-Paar zugeordneten Leistungswert jeweils einzeln errechnen. Die Leistungsbewertung von RS mittels Objektexperimenten ist ganz offensichtlich eine sehr aufwendige Technik. Sie besitzt darüber hinaus den zusätzlichen Nachteil, daß sie während Entwurfsphasen prinzipiell nicht einsetzbar ist (Entwurf bedeutet ja, daß das Objekt noch nicht existiert). Wir besinnen uns daher auf die in Kap. 1 bereits skizzierte, hoffnungsvolle Alternative: Anstatt ein Objekt-System direkt zu beobachten, versuchen wir, ein Ersatz-System zu schaffen, das einerseits dem zu untersuchenden Objekt in den wesentlichen Eigenschaften "stark ähnelt", andererseits aber "leichter manipulierbar" ist als das Objekt selbst. Jedes derartige Ersatzsystem hatten wir "Modell" getauft; uns geht es speziell um Modelle, die in der Lage sind, unsere Leistungsfunktion (s. oben) in konkreten Fällen anstelle der Objektwelt und in hinreichender Ähnlichkeit zu ihr zu repräsentieren. Solche Modelle nennen wir "Leistungsmodelle".

Modelle

mathematische Modelle

Wir wollen uns in dieser Vorlesung ausschließlich mit mathematischen Modellen befassen. Bei diesem Modelltyp geht es darum, aus Angaben über Struktur und Parameterwerte von System und Last, (zugehörige) Leistungswerte mittels mathematischer Techniken zu "errechnen", bzw. einen mathematischen Zusammenhang der diversen involvierten Größen zu beschreiben. Als Idealziel eines solchen Unterfangens wären zweifellos geschlossene Formeln anzusehen, die ein leichtes Hantieren in einer Art erlauben, wie sie etwa bei dem bekannten mathematischen Modell des freien Falls $s = g/2 \cdot t^2$ vorliegt. Wir werden dieses Ziel nicht in dieser idealen Form und auch nicht für alle denkbaren Problemfälle erreichen. Wir werden aber für eine praktisch relevante Klasse von Problemen implizite mathematische Zusammenhänge zwischen Maschinenparametern, Lastparametern und Leistungswerten ableiten, die es (im Rahmen dieser Problemklasse) erlauben, Leistungswerte auf mathematischem Wege zu ermitteln.

Lernziele

Dazu werden wir im vorliegenden Kapitel (als Lernziel)

Verkehrsnetze

- eine strukturierte Vorstellungswelt für mathematische Leistungsmodelle, die sog. Verkehrsnetze, kennenlernen sowie die Fähigkeit erwerben, Probleme der realen Welt in diese Modelwelt zu übersetzen bzw. zu erkennen, wo eine solche Übertragung scheitert;

Betriebs-analyse

- in einer ersten quantitativen Interpretation der Verkehrsnetze eine auf der Vorstellung von Messungen aufbauende Analysetechnik, die sog. Betriebsanalyse, betrachten; wir werden dabei eine Reihe, auch für zukünftige Überlegungen wesentlicher, Größen und Begriffe definieren (so: Ankunfts- und Abgangsrate; Bedienbedarf, Bediengeschwindigkeit und Bedienzeit; arbeitserhaltende Station; Übergangsrate und Übergangshäufigkeit; Auslastung und Flaschenhals) und mit ihnen umzugehen lernen; wir werden erkennen, daß diverse Meßgrößen formelmäßig zusammenhängen und werden diese Zusammenhänge (konkret: Auslastungsgesetz, Verkehrsflußgleichgewicht, Flaschenhals beim offenen und geschlossenen Modell) einzusetzen lernen;

stochastische Verkehrsnetze

- in einer weiteren, stochastischen Interpretation der Verkehrsnetze diese soweit konkretisieren, daß sie als Generatoren stochastischer Betriebsabläufe erkennbar sind; wir werden verstehen lernen, daß die mathematische Modellklasse "stochastischer Prozeß" zur Analyse der Zusammenhänge zwischen den Eigenschaften eines stochastischen Verkehrsnetzes einerseits und den Merkmalen der Menge generierbarer Betriebsabläufe andererseits geeignet ist; wir werden lernen, mit dem Begriff "Zustand" umzugehen und daraus einen prinzipiellen Plan für die (in den folgenden Kapiteln vorzunehmende) effektive Analyse stochastischer Verkehrsnetze ableiten.

analytische Modelle

In der Literatur stößt man häufig auf den Begriff "analytische Modelle". Dabei sind meist die im letzten Absatz skizzierten mathematischen Modelle auf der Basis stochastischer Verkehrsnetze gemeint.

2.1 Modellbildung: Verkehrsnetze

Vielen gebräuchlichen Leistungsmodellen des mathematischen Typs liegt eine Abstraktion struktureller Natur zugrunde, nach der ein Rechensystem in der Form eines Verkehrsnetzes dargestellt und untersucht wird. Ein Verkehrsnetz besteht (in der hier verwendeten Terminologie) aus zwei Komponenten, der "Maschine" und der "Last". Die Maschine wird dabei durch einen (zeitlich festen) gerichteten Graphen repräsentiert, die Last durch eine (oft zeitlich variierende) Menge von Prozessen.

Verkehrsnetz

Werden wir konkreter: Die Maschine besteht aus einer Menge von "Betriebsmitteln" (auch: "Ressourcen", "Funktionseinheiten", "Stationen") sowie einer Menge von "Übergangsmöglichkeiten" (auch: "Verbindungen", "Wegen") zwischen gewissen Paaren dieser Betriebsmittel. In graphischer Darstellung erhalten wir somit den erwähnten Graphen, ein Netz aus Stationen:

Maschine

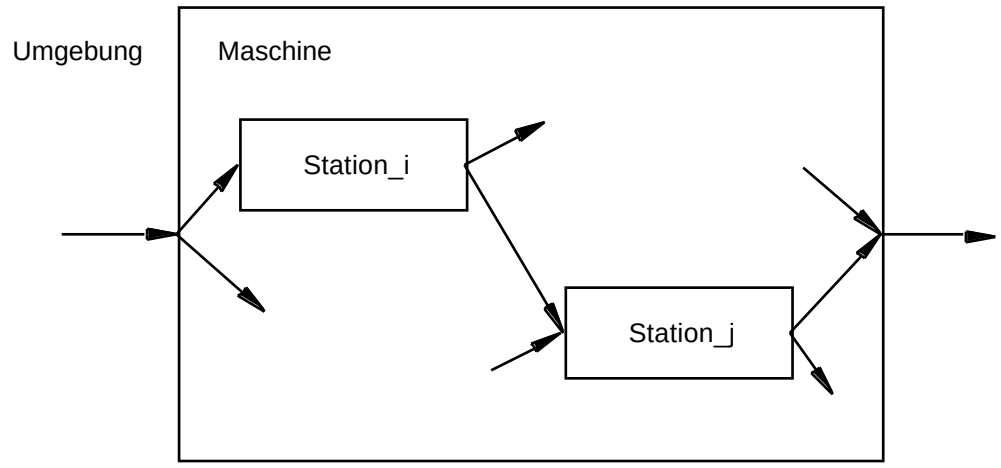


Abbildung 2.1.1: Maschine eines Verkehrsnetzes

Die obige schematische Darstellung gibt exemplarisch nur zwei Betriebsmittel wieder (genannt "Station_i" und "Station_j") - es könnten natürlich mehr sein; auch ist exemplarisch nur eine Übergangsmöglichkeit (zwischen Station_i und Station_j) eingezeichnet - es könnten natürlich mehr sein, so z.B. von Station_i nach anderen Betriebsmitteln, von anderen Betriebsmitteln nach Station_j, u.s.w. Des weiteren unterscheidet die Darstellung explizit zwischen der Maschine und ihrer Umgebung (dem "Rest der Welt") und weist Übergangsmöglichkeiten aus der Umgebung in die Maschine (zu gewissen Betriebsmitteln) und aus der Maschine (von gewissen Betriebsmitteln) in die Umgebung aus.

Die zweite Komponente, die Last, besteht aus einer Menge von "Prozessen" (je nach Auffassung auch: "Lasteinheiten", "Aufträgen", "Tasks", "Jobs", "Kunden"). Jeden Prozeß stelle man sich zunächst statisch vor als eine zusammengehörige Menge von Betriebsmittelanforderungen (Einzelanforderungen an einzelne Betriebsmittel der Maschine) und eine Reihenfolgevorschrift, die eine zeitliche (u.U. nur partielle) Ordnung über den Betriebsmittelanforderungen festlegt. Wir werden zur Erhöhung der Verständlichkeit die Begriffe Betriebsmittelanforderung und Reihenfolgevorschrift später detaillierter diskutieren. Zunächst aber: Wir haben in dieser statischen Charakterisierung eines Prozesses eine Verhal-

Last

tensvorschrift vor uns, die wir "Prozeßmuster" nennen wollen. Ein Prozeß im dynamischen Sinne entsteht, wenn die Maschine dieses Prozeßmuster als Arbeitsauftrag versteht, dem sie Schritt für Schritt nachkommt, indem sie Betriebsmittelanforderung nach Betriebsmittelanforderung in der vom Prozeßmuster geforderten Abfolge erfüllt. Bleibt zu klären, wie Prozesse (in unserer Modellwelt) entstehen, wofür wir zwei unterschiedliche Fälle vorsehen:

temporäre Prozesse

- Im einen Fall generiert eine nicht näher betrachtete Umgebung den Prozeß und übergibt ihn der Maschine, von welchem Zeitpunkt ab er existiert; wir können uns z.B. den "Benutzer" vorstellen, der sein Programm irgendwie "startet"; in diesem Fall ist es auch sinnvoll, anzunehmen, daß der Prozeß irgendwann (nach vollständiger Abarbeitung) wieder terminiert; wir nennen einen solchen Prozeß "temporär" und ein System/Modell, das nur temporäre Prozesse vorsieht, ein "offenes".

permanente Prozesse

- Im anderen Fall wird der Prozeß weder generiert noch terminiert; er existiert "dauernd" (und muß daher in seiner Anforderungsreihenfolge wohl irgendwie zyklisch sein); wir nennen einen solchen Prozeß "permanent" und ein System/Modell, das nur permanente Prozesse vorsieht, ein "geschlossenes". Diese (auf den ersten Blick sicher wirklichkeitsfremd anmutende) Vorstellung permanenter Prozesse wird sich schnell als äußerst hilfreiche Fiktion erweisen.

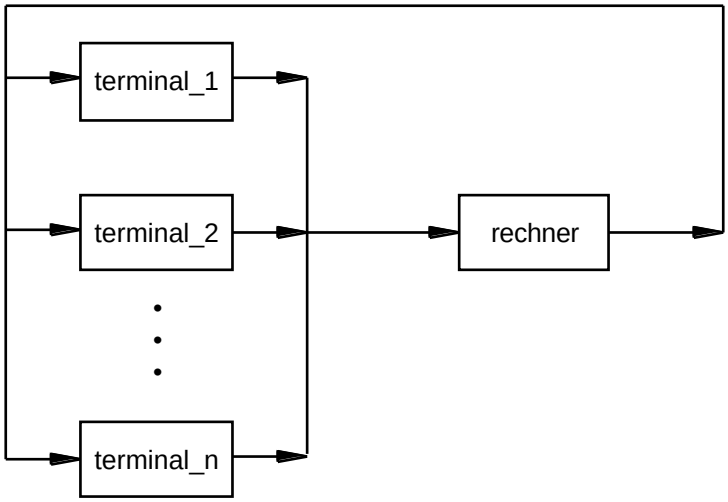
Testfrage

Testfrage 2.1.2: Warum ist es i.allg. sinnvoll, anzunehmen, daß ein explizit generierter ("der Maschine übergebener") Prozeß auch wieder terminiert?

Schieben wir hier einige veranschaulichende Überlegungen ein. Die skizzierte Vorstellung von Verkehrsnetzen wird häufig ohne detailliertere Erklärung in Vorlesungen und Büchern über Rechnerarchitektur, Betriebssysteme u.ä. eingesetzt. Wir greifen willkürlich zwei typische Beispiele heraus.

Beispiel

Beispiel 2.1.3: Häufig stößt man auf folgendes Verkehrsnetz als simples Modell eines Teilnehmerrechensystems. Dabei wird die Maschine (in unserer jetzigen Terminologie) durch einen Graphen repräsentiert:



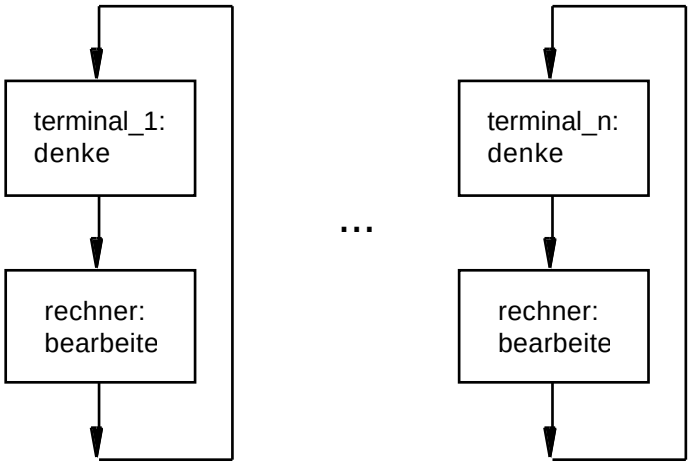
Wir haben es hier mit n+1 Stationen zu tun: Den Endgeräten "terminal_1" bis "terminal_n" und dem eigentlichen "rechner". Als Wege sind vorgesehen die

Übergangsmöglichkeiten von terminal_i nach rechner und von rechner nach terminal_i, (i=1,2,...,n). Das Modell hat keine Ein/Ausgänge von/nach irgendeiner Umgebung, ist also geschlossen.

Von der Last besteht ungefähr folgende Vorstellung: In jeder Station terminal_i (i=1,2,...,n) wird eine Weile nachgedacht (und dabei wohl ein Auftrag vorbereitet), dann wird ein Auftrag an die Station rechner gesandt (die diesen dann wohl bearbeitet), dann wird wieder nachgedacht, dann bearbeitet u.s.w. Diese verbal geschilderten Verhaltensregelmäßigkeiten sind in unserem Sinne "Prozeßmuster". Über eine formalere Notation für Prozeßmuster haben wir uns noch keine Gedanken gemacht; naheliegend und einleuchtend sind folgende Formulierungen,

Notation für Prozeßmuster

- eine graphische Notation:



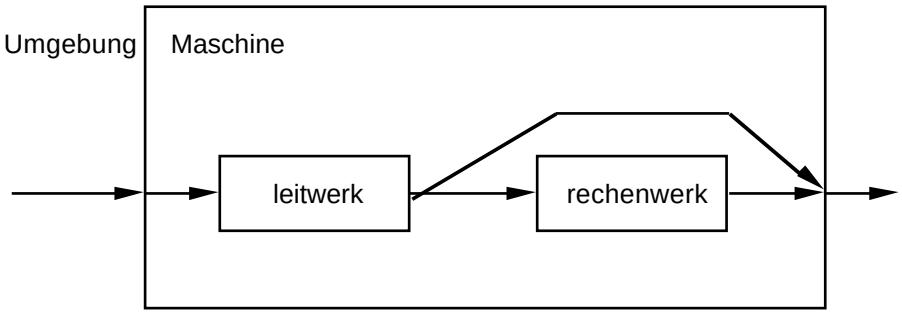
- eine sprachliche Notation:

```
LOOP                              LOOP
   terminal_1.denke;                 terminal_n.denke;
   rechner.bearbeite     ...         rechner.bearbeite
ENDLOOP                           ENDLOOP
```

beide das Wesentliche ausdrückend: In fortwährender Wiederholung wird von einem terminal_i eine "denke"-Tätigkeit verlangt, anschließend von rechner eine "bearbeite"-Tätigkeit u.s.w. Wir haben es insgesamt mit n Prozeßmustern zu tun (jedes ist den Vorgängen bzgl. genau einer Station_i zugeordnet); um die Beschreibung zu komplettieren, können wir sinnvollerweise (an jedem Terminal sitzt wohl nur ein Benutzer) festlegen, daß es zu jedem Prozeßmuster genau einen Prozeß gibt. Alle Prozesse sind permanent (wir machen uns ganz einfach keinerlei Gedanken darüber, wie der Betrieb beginnt oder endet), wie auch die Prozeßmuster ausweisen.

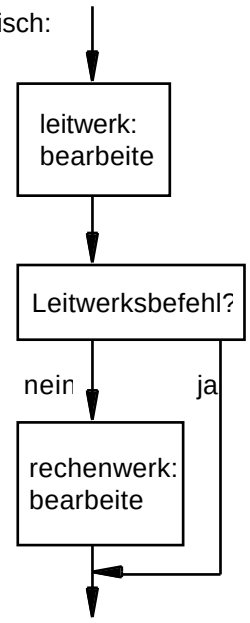
Beispiel

Beispiel 2.1.4: Betrachten wir folgendes Verkehrsnetz als simples Modell eines Zentralprozessors. Als Maschine diene der umseitig slizzierte Graph. Zwei Stationen also: "leitwerk" und "rechenwerk"; dazu die Wege von Umgebung nach leitwerk, von leitwerk nach rechenwerk und Umgebung, von rechenwerk nach Umgebung. Das Modell hat Ein- und Ausgang, ist also offen.



Die Last unterliegt folgender Vorstellung: Aufträge betreten das Netz, leitwerk muß zunächst für jeden Auftrag tätig werden, danach ist entweder die Bearbeitung bereits abgeschlossen (im Falle eines "Leitwerksbefehls") oder rechenwerk muß tätig werden (falls kein reiner Leitwerksbefehl vorliegt). Die Entscheidung Leitwerksbefehl bzw. kein Leitwerksbefehl erfolgt (im Modell!) zufällig, mit Wahrscheinlichkeiten 1-r bzw.r. Dabei machen wir uns im jetzigen Kontext über die Darstellung des "Würfelns" Leitwerksbefehl/kein Leitwerksbefehl keine Gedanken. Als Prozeßmuster erhalten wir:

Graphisch:



Sprachlich:

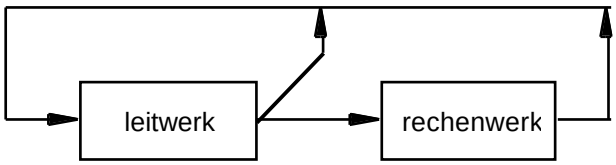
```
leitwerk.bearbeite;
IF  "kein Leitwerksbefehl"
THEN  rechenwerk.bearbeite;
```

Wir haben es mit genau einem Prozeßmuster zu tun, dem jeder generierte Prozeß folgt; es kann (im Lauf der Zeit) beliebig viele solche Prozesse geben, alle sind temporär. Um die Beschreibung zu komplettieren, müßten wir festlegen, wann die diversen einzelnen Prozesse generiert werden sollen (in der Terminologie des Beispiels: wann die einzelnen Befehle dem Zentralprozessor zur Ausführung übergeben werden). Bevor wir uns in der Erfindung eines diesbezüglichen "Generator-Prozesses" verlieren (im allg. ist ein solcher sicher vonnöten und könnte im einfachsten Fall auf einer Tafel mit Generierungszeitpunkten basieren), sollten wir beachten, daß beliebig festgelegte Prozeß-Generierungszeitpunkte dem aktuellen Beispiel nicht gerecht werden: Der Zentralprozessor bearbeitet ja ein Programm, das im Normalfalle sofort nach Abarbeitung eines Befehls einen nächsten zur Abarbeitung bereithält. Wenn wir den Fall keines verfügbaren Programs ganz einfach ignorieren, dürfte folgendes Verkehrsnetz die

Sachlage besser treffen:

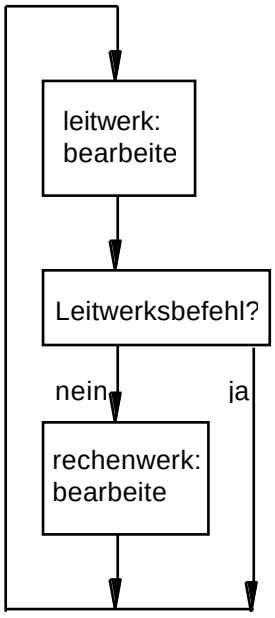
Beispiel in alternativer Form

Maschine:



Bemerke: Das Modell ist geschlossen

Last (Prozeßmuster):




```
LOOP
    leitwerk.bearbeite
    IF "kein Leitwerksbefehl"
    THEN rechenwerk.bearbeite
ENDLOOP
```

Last (Prozesse):

Es gibt genau einen permanenten Prozeß zu obigem Prozeßmuster, der ein ganzes Programm, mehr noch: eine Folge von Programmen, präziser: die Aufträge einer Folge von Programmen an einen Zentralprozessor, repräsentiert.

Kehren wir nach diesen Beispielen zur Hauptlinie zurück, für die eine detailliertere Diskussion der Begriffe Betriebsmittelanforderung und Reihenfolgevorschrift noch ausstand.

Zu den Betriebsmittelanforderungen:
Ohne dies explizit zu erwähnen, hatten wir in den Beispielen Anforderungen an die Stationen (Betriebsmittel) gestellt, die funktionaler Natur waren. So sollte in Bsp. 2.1.3 "terminal_i" die Tätigkeit "denke", "rechner" die Tätigkeit "bearbeite"

vollziehen u.s.w. Etwas konkreter ausgedrückt war unsere implizite Vorstellung, daß

- Stationen gewisse benannte "Dienste" zu erbringen fähig waren,
- Prozesse die Erbringung dieser Dienste von den Stationen fordern konnten.

Diese Darstellung erinnert Sie sicher (und soll dies auch tun) an Wissensbereiche, die Ihnen von anderen Vorlesungen her geläufig sind, wie z.B. "Abstrakte Datentypen" oder "Objekt-orientierte Programmierung". Im Unterschied zu diesen Bereichen sind wir aber hier (ich erinnere an Kap. 1)

- an den informationsorientierten Wirkungen der Erbringung von Diensten (also den Resultaten von Berechnungen oder den Veränderungen von Speicherinhalten) wenig interessiert;
- an den physikalischen Wirkungen der Erbringung von Diensten (also der Dauer der Beanspruchung von Prozessoren oder des Umfangs der Beanspruchung von Speichern) stark interessiert.

Betriebsmittelanforderungen

Es ist daher für Leistungsmodelle (insbesondere auch für die hier betrachteten Verkehrsnetze) bequemer und allgemein gebräuchlich, Betriebsmittelansprüche nicht funktional, sondern in physikalischen Begriffen (Zeit und Raum) anzugeben.

So etwa in Bsp. 2.1.3

nicht:	terminal_i.denke
sondern:	terminal_i."widme mir (z.B.) 20 sec"

Oder für eine CPU

nicht:	CPU.bearbeite
sondern:	CPU."widme mir (z.B.) 1 μsec"

Der Usus, die Anforderungen an Betriebsmittel direkt durch deren physikalische Beanspruchung auszudrücken, bringt allerdings auch Nachteile mit sich. Der wohl gravierendste besteht in der Tatsache, daß eine Angabe der physikalischen Beanspruchung eines Betriebsmittels nur möglich ist, wenn Interna der Bearbeitung im Betriebsmittel (und damit in der modellierenden Station) bekannt sind. Um dies wieder am Beispiel klarzumachen: Die Übersetzung (s. Bsp. 2.1.3)

von:	rechner.bearbeite
nach:	rechner."widme mir ??"

ist angesichts der anzunehmenden internen Komplexität von "rechner" nicht ohne weiteres möglich; zumindest nicht ohne Interna von "rechner" zu kennen; und auch dann voraussichtlich nicht als simple eindimensionale Größe (wie sich das etwa bei "Terminals" und "CPUs" anbietet).

Stationen

Als Folge dieser Beobachtung sehen wir uns genötigt, in unseren Verkehrsnetzen die Modellkomponente "Station" nicht als Abbild eines beliebig komplexen "Subsystems" der realen Welt zu verwenden; vielmehr sehen wir Stationen nur als Repräsentanten relativ einfacher Subsysteme, wobei "relativ einfach" dadurch definiert sei, daß wir berechtigt sind (oder uns berechtigt fühlen), in etwa folgende interne Struktur einer Station zu postulieren

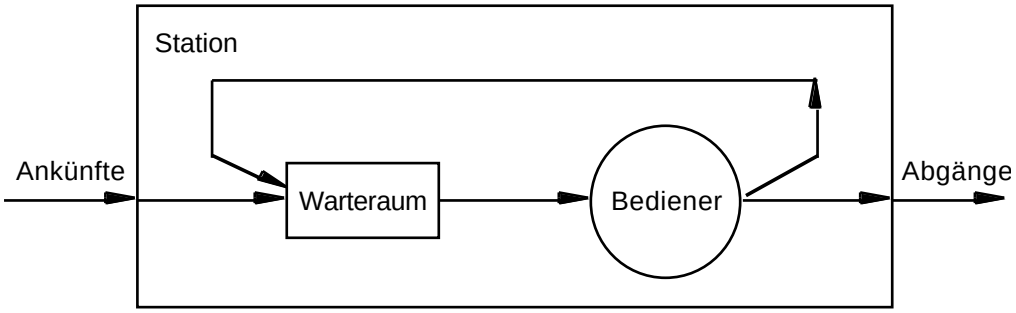


Abbildung 2.1.5: Struktur einer Bedienstation

Ankommende Kunden (Ankünfte) betreten die Station mit einem Bedienwunsch, der eine (eindimensionale) physikalische Beanspruchung des Bedieners ausdrückt. Die Bedienstation widmet sich diesem Bedienwunsch nach eigenen internen Regeln: Sie kann den Kunden z.B. warten lassen, seine Bedienung unterbrechen u.ä. mehr. Schließlich, nach Abarbeitung seines Bedienwunsches, wird der Kunde entlassen (Abgänge). Bei den Stationen und den für sie möglichen Bedienwünschen unterscheidet man zwei wesensverschiedene Arten:

aktive Station

- die "aktive" Bedienstation, die die zeitliche Belegung eines "Prozessors" verwaltet und mit Bedienwünschen der Dimension "Zeiteinheit" konfrontiert wird; Beispiele sind Stationen, die eine Zentraleinheit, einen Kanal, eine Übertragungsleitung u.ä. repräsentieren;

passive Station

- die "passive" Bedienstation, die die räumliche Belegung eines "Speichers" verwaltet und mit Bedienwünschen der Dimension "Raumeinheit" konfrontiert wird; Beispiele sind Stationen, die einen Arbeitsspeicher, einen Pufferbereich u.ä. repräsentieren.

Wir werden allerdings die Berücksichtigung räumlicher Bedienwünsche in mathematischen Modellen bald wieder fallenlassen müssen (leider!).

Zu den Reihenfolgevorschriften:

Reihenfolgevorschriften

Wir hatten akzeptiert, daß zwischen den einzelnen Betriebsmittelanforderungen eines Last-Prozesses Reihenfolgevorschriften bestehen können etwa der Art: "Erst muß Tätigkeit a abgeschlossen sein, dann kann Tätigkeit b begonnen werden". Wir hatten in den Beispielen 2.1.3 und 2.1.4 bei den versuchsweisen Notationen für Prozeßmuster solche Reihenfolgevorschriften auch bereits freizügig ausgedrückt (s. dort). Auch tritt Ihnen die Notwendigkeit, Reihenfolgen zwischen einzelnen Arbeitsschritten eines Prozesses zu berücksichtigen, nicht zum ersten Mal entgegen. So werden im Bereich Betriebssysteme u.a. Präzedenzgraphen zur Festlegung der Ordnung über Teilschritten nicht-zyklischer Prozesse eingesetzt. Ich erwähne dies deshalb, weil bei diesen Präzedenzgraphen besonders deutlich wird, daß die Ordnung i.allg. nicht total, sondern nur partiell ist. Ein Lastprozeß könnte durchaus (von sich aus) zwischen gewissen seiner Teilschritte keine strikte Reihenfolge vorsehen:

Präzedenzgraph

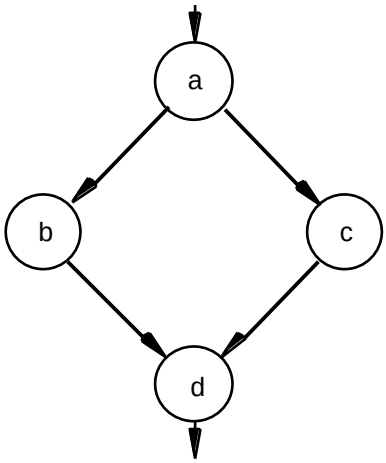


Abbildung 2.1.6: Präzedenzgraph mit partieller Ordnung

Das einfache Beispiel der Abb. 2.1.6 drückt aus: a muß vor b und c erledigt werden, d nach b und c, implizit damit auch d nach a; andererseits ist zwischen b und c keine Reihenfolge festgelegt: b könnte (vom Prozeß aus gesehen) vor oder nach c "drankommen"; b und c könnten auch gleichzeitig "laufen" oder irgendwie "zeitlich überlappend" u.s.f.

gleichzeitige Belegung von Betriebsmitteln Zusammenfassung

Die erst in späteren Kapiteln deutlich werdende Schwierigkeit der analytischen Behandlung von Leistungsmodellen bringt es mit sich, daß wir den Fall der gleichzeitigen Belegung mehrerer Betriebsmittel durch einen Prozeß nicht in voller Allgemeinheit berücksichtigen können. Wir bescheiden uns daher schon jetzt und fordern für die Reihenfolgevorschriften eine totale Ordnung - wodurch aus den Lastprozessen sequentielle Prozesse werden, aus den Prozeßmustern so etwas wie sequentielle Programme. Natürlich reduziert das unsere Möglichkeiten, die Realität adäquat zu modellieren: Die schmerzlichste Folgerung ist die, daß wir die passiven Bedienstationen wieder vergessen müssen: "Speicherplatz" wird nahezu immer gleichzeitig mit "Prozessorzeit" benötigt - gleichzeitige Belegung von Betriebsmitteln wollten wir aber ausschließen.

Fassen wir zusammen: Strukturell gesehen, stellen Verkehrsnetze eine geeignete Modellklasse zur Beschreibung von Leistungsmodellen dar. Ein Verkehrsnetz setzt sich zusammen aus einer Maschine und einer Last. Die Maschine besteht aus einer Menge von Stationen samt Übergangsmöglichkeiten zwischen den Stationen, wobei die Stationen typmäßig bis auf weiteres auf den Bereich der aktiven Bedienstationen beschränkt sind. Die Last wird beschrieben durch eine Menge von Prozeßmustern sowie Regeln, welche die Generierung bzw. Zahl von Prozessen bezüglich einzelner Prozeßmuster festlegen. Jedes Prozeßmuster notiert alle Bedienwünsche zugehöriger Prozesse und ihre Reihenfolge, wobei bis auf weiteres die Bedienwünsche als zeitliche Beanspruchung jeweils eines Bedieners (einer Station) angegeben werden und die Bedienwünsche streng sequentiell aufeinanderfolgen. Der Verkehr im Netz (als Abbild des Betriebs des Rechensystems) entsteht durch die Bewältigung der Bedienwünsche aller Lastprozesse seitens der Stationen der Maschine, wobei dank vorgegebener Restriktionen jeder Lastprozeß dynamisch als "Kunde" gesehen werden kann, der von Station zu Station wandert (insbesondere zu jedem Zeitpunkt in genau einer Station verweilt).

2.2 Betriebsanalyse (Operational Analysis)

Einordnung

Das Konzept der Betriebsanalyse (oft auch als "operationale Analyse" ins Deutsche übersetzt) ist in einer Reihe von Arbeiten eingeführt und angewendet worden (guter Übersichtsartikel: DeBu78; wichtige Erweiterung: Rood79, durchgängige Verwendung im Lehrbuch LZGS84). Obzwar im allgemeinen nicht zu den analytischen Modellen gerechnet, verwenden die Modelle der Betriebsanalyse unzweifelhaft eine Technik, die auf mathematischem Weg Zusammenhänge zwischen Kenngrößen von Maschine, Last und Leistung aufdeckt: Die Betriebsanalyse zählt daher in unserem Sinne zu den mathematischen Modellierungstechniken. Vor allem aber ist die Betriebsanalyse eine Technik, die den Vorgang der Modellbildung im Bereich analytischer Leistungsmodelle auf sehr leicht faßbare Weise unterstützt.

Meßbarkeit, Überprüfbarkeit

Es zählt zu den Prinzipien der Betriebsanalyse, daß alle verwendeten Größen während des Betriebs eines Rechensystems meßbar sind (bzw. aus solchen meßbaren Größen mittels klar definierter Formeln errechnet werden können) sowie daß sämtliche einfließenden Annahmen während des Betriebs eines Rechensystems mittels Messungen überprüft werden können (daher auch der Name der Technik: "Betriebs"-Analyse, "operational" analysis). Um evtl. Mißverständnissen von vornherein vorzubeugen: Wir setzen nicht voraus, daß alle verwendeten Größen tatsächlich während des Realbetriebs gemessen werden - sie sollten nur "im Prinzip" meßbar sein.

Meßpunkte

Zur Definition der Meßpunkte bemühen wir die vorgestellte Maschinenstruktur nach Abb. 2.1.1, sehen das Rechensystem also von vornherein als Verkehrsnetz. Die Meßpunkte liegen jeweils an Stations-Eingängen und -Ausgängen, so daß wir den Verkehr der Kunden in die Stationen hinein und aus den Stationen heraus beobachten können. Nicht unmittelbar beobachtbar sind damit natürlich die Prozeßmuster (die "treibende Kraft" des Betriebs).

Station

Konzentrieren wir uns zunächst auf eine beliebige einzelne Station des Verkehrsnetzes:

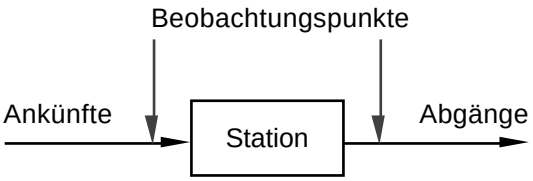


Abbildung 2.2.1: Einzelstation

Wir halten folgende (im Prinzip sicher meßbaren) Größen fest:

Basisgrößen

Definition 2.2.2: Basisgrößen

T Gesamtdauer der Beobachtung (Länge Beobachtungsintervall)

sowie die während des Beobachtungsintervalls beobachteten

A Zahl der Kundenankünfte ("arrivals")

C Zahl der Kundenabgänge ("completions")

B Belegtzeit der Station ("busy time"),
definiert als Gesamtzeit (Summe aller Zeitintervalle) mit mindestens einem Kunden an der Station

Als Anmerkung: Zur Messung von B ist es erforderlich, zu jedem Zeitpunkt den Belegungszustand der Station zu kennen ("wieviele Kunden sind drin?") - zumindest zu erkennen, ob überhaupt ein Kunde anwesend ist. Der Belegungszustand läßt sich aber bei Kenntnis des Anfangszustands (zu Beginn der Beobachtung, z.B. "leer") aus der Zahl der Ankünfte und Abgänge "von außen" ermitteln.

Aus den Meßgrößen lassen sich leicht folgende Größen ableiten:

Abgeleitete Größen

Definition 2.2.3: Abgeleitete Größen

$a := A/T$
$c := C/T$
$b := B/T$
$\overline{B} := B/C$

Benennungen

Für die rein formal abgeleiteten Größen der Def. 2.2.3 haben sich bestimmte Benennungen eingebürgert. Diese legen allerdings Interpretationen nahe, die nicht in allen Fällen korrekt sind und deshalb genau diskutiert werden sollen. Zunächst spielt bei allen Benennungen die Vorstellung einer relativ großen Beobachtungszeit T und einer gewissen statistischen Regelmäßigkeit eine Rolle. Unter diesen Voraussetzungen sind die Benennungen

a (=A/T) "Ankunftsrate"
c (=C/T) "Abgangsrate"
b (=B/T) "Belegungsgrad", "Belegungsanteil"

einleuchtend und hilfreich.

Testfrage

Testfrage 2.2.4: Inwiefern sind die Vorstellungen einer "relativ großen Beobachtungszeit" und einer "gewissen statistischen Regelmäßigkeit" wesentlich für die Angemessenheit der Benennungen Ankunfts-, Abgangs-"Rate"?

Mittlere Bedienzeit

Wesentlich vorsichtiger zu genießen ist dagegen die Benennung

$\overline{B}$ (=B/C) "mittlere Bedienzeit"

Erinnern wir uns an unsere prinzipiellen Annahmen über Verkehrsnetze: Kunden betreten eine Station mit einem Bedienwunsch, dessen Größe wir durch die Dauer der (bevorstehenden) Beanspruchung des Bedieners angeben wollten. Über eine gewisse Menge von Kunden (mit in der Regel unterschiedlichen Bedienwünschen) gemittelt, könnten wir einen mittleren Bedienwunsch errechnen, der nach unserer Konvention der mittleren zeitlichen Beanspruchung des Bedieners pro Kunde entspricht. Nun sind uns in der Betriebsanalyse (aufgrund der festgelegten Meßtechnik) die Bedienwünsche der Kunden nicht direkt zugänglich, so daß wir versucht sein werden, Kenngrößen für Bedienwünsche (z.B. deren Mittel) aus den (gemessenen) Basisgrößen zu ermitteln. Die Schlußfolgerung aber

Mittlerer Bedienwunsch pro Kunde
= mittlere zeitliche Beanspruchung des Bedieners pro Kunde
(aufgrund der Konvention der Angabe)
= "mittlere Bedienzeit"
(auch diese Definition wäre sinnvoll - und wird auch benutzt!)
= $\overline{B}$

Doppeldeutigkeit

ist nur unter zusätzlichen Annahmen über die Funktionsweise der Bedienstation korrekt (und deckt damit die Gefährlichkeit der in zweifachem Sinne gebrauchten Benennung "mittlere Bedienzeit" auf).

Gehen wir dieser Frage etwas weiter nach. Habe unsere aktive Bedienstation die einfache Struktur der Abb. 2.1.5. Wie "bedient" die Station die ankommenden

Bedienungsmechanismus

Kunden? Nehmen wir zunächst an, die Station sei bei Eintreffen eines Kunden unbeschäftigt (englisch: "idle") und habe auch genügend Zeit, seine Bedienung vor Eintreffen des nächsten Kunden zu beenden. Wir lockern unsere Konvention der Angabe von Bedienwünschen etwas, indem wir akzeptieren, daß der Bedienwunsch des Ankömmlings zunächst in irgendwelchen eindimensionalen "Arbeitseinheiten" (AE) angegeben wird (Zahl auszuführender Instruktionen, Länge zu übertragender Nachricht o.ä.), wobei der konkrete Bedienwunsch des betrachteten Neuankömmlings die Größe W habe. Der Bediener in der Station wende sich unmittelbar der Bedienung des Neuankömmlings zu und führe diese ohne Unterbrechung zu Ende. Er benötigt dazu ("aktive" Station!) eine gewisse Zeit S (die "Bedienzeit"), die bei sinnvoller Wahl der Maßeinheit AE monoton wachsend mit W zusammenhängt. Wir lassen ausschließlich eine direkte Proportionalität zwischen W und S zu, in Zeichen

(2.2.5) $S = W / r$

Bearbeitungsgeschwindigkeit

so daß r auch als "Bearbeitungsgeschwindigkeit" interpretiert werden kann: Die Station schafft r AE je ZE (Zeiteinheit). Der einfachste Fall ist offensichtlich r=1; kein spezieller Sonderfall, sondern in unserem Belieben, wenn wir uns entschließen, Bedienwünsche in Bezug auf die speziellen Stationsfähigkeiten zu messen: "Der Bedienwunsch entspricht einer Bedienzeit von S Zeiteinheiten" bezeichnet genau unsere Konvention, Bedienwünsche durch Beanspruchungen des Bedieners zu kennzeichnen.

Bediendisziplin

Wie bedient die Station in komplizierteren Fällen? Stellen wir uns die Organisation des Betriebs innerhalb der Station geregelt vor durch eine "Bediendisziplin". Diese hält einen Teil der anwesenden Kunden im Warteraum (wo sie einfach warten, also nicht bedient werden), den Rest der anwesenden Kunden in der Bedieneinrichtung (wo sie alle gleichzeitig einen Fortschritt in ihrer Bedienung erfahren). Der einfachste, leicht vorstellbare Fall ist offensichtlich die Bedienung genau eines Kunden zu jedem Zeitpunkt (wenn die Station nicht "leer" ist); andere Fälle werden wir als bequeme Modell-Fiktionen später kennenlernen. Nehmen wir nun zusätzlich an, die Bearbeitungsgeschwindigkeit des Bedieners sei konstant gleich r, gleichgültig wieviele Kunden in der Station weilen (in diesem Kontext verwendet man gerne auch die Bezeichnung "Arbeitskapazität" statt "Bearbeitungsgeschwindigkeit") und diese Arbeitskapazität komme voll den anwesenden Kunden zugute (der Bediener legt also insbesondere keine "Pausen" ein, wenn Arbeit vorhanden ist und er hat auch keine sonstigen "Verwaltungstätigkeiten", englisch: "overhead activities", zu leisten). Abb. 2.2.6 stellt einen möglichen Verlauf des Stationszustandes über der Zeit dar (ein "Belegungsgebirge").

Arbeitskapazität

Wir beobachten C (in der Abbildung: 5) Abgänge. Seien die Bedienwünsche der zu diesen C Abgängen gehörigen Kunden (in irgendeiner Reihenfolge) mit $W_1, W_2, ..., W_C$ bezeichnet. Die gesamte zu leistende Arbeit betrug also

(2.2.7a) $$W_{total} = \sum_{i=1}^{C} W_i$$

Jedem Bedienwunsch W_i entspricht (Annahme: isolierte, ungestörte Bearbeitung) nach (2.2 .5) eine Bedienzeit

(2.2.7b) $S_i = W_i / r \qquad i=1,2,\ldots,C$

Belegungs-gebirge

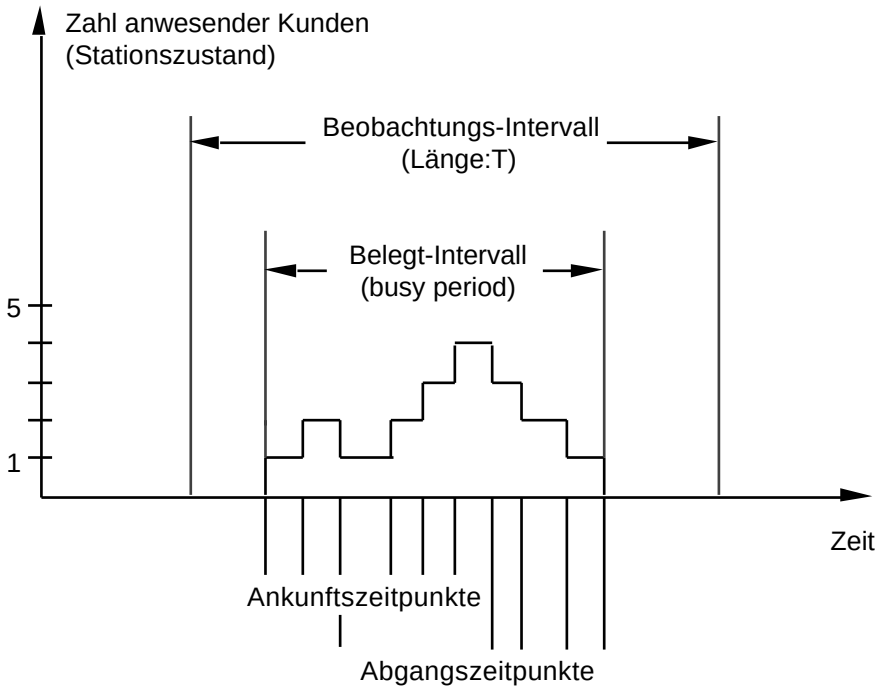


Abbildung 2.2.6: Belegungsgebirge

Als mittlere Bedienzeit (im Sinne der mittleren Belastung des Bedieners pro Kunde) erhalten wir für diese C Kunden

(2.2.7c) $$\bar{S} = \frac{1}{C} \sum_{i=1}^{C} S_i = \frac{1}{C\,r} \sum_{i=1}^{C} W_i$$

und bei lückenlosem "Aneinanderlegen" aller Bedienzeiten eine totale Bedienzeit

(2.2.7d) $$S_{total} = \sum_{i=1}^{C} S_i = \frac{1}{r} \sum_{i=1}^{C} W_i = C\ \bar{S}$$

Andererseits benötigt der Bediener zur Abarbeitung der gesamten Arbeit W_{total} (wegen der konstanten Geschwindigkeit r) die Zeit

(2.2.7e) $$B = W_{total}/r$$

die wir schon mit B bezeichnet haben, da sie genau der Länge des Belegt-Intervalls (und damit der Belegtzeit gemäß Def. 2.2.2) entspricht: Der Bediener ist ja so lange tätig, wie Arbeit vorhanden ist - ein Belegt-Intervall beginnt mit der Ankunft eines "ersten" Kunden (bei leerer Station) und endet mit der erfolgten Bewältigung der gesamten Arbeit dieses ersten Kunden und derer, die nachfolgend (bei nicht-leerer Station) eintrafen. Aus (2.2.3, 2.2.7d, 2.2.7e) erhält man leicht

(2.2.7f) $$S_{total} = B \quad \text{bzw.:} \quad \bar{S} = \bar{B}$$

womit in diesem Fall die "mittlere Bedienzeit" $\bar{B}$ im Sinne der Def. 2.2.3 und die "mittlere Bedienzeit" $\bar{S}$ im Sinne von (2.2.7c) tatsächlich zusammenfallen. Wie man sich leicht überlegen kann, bleibt die Gleichheit erhalten, solange die Bedienstation "arbeitserhaltend" und "von konstanter Bedienkapazität" ist:

arbeits-erhaltende Station

Definition 2.2.8: Eine arbeitserhaltende Bedienstation (vom Englischen "work-conservative", kurz "conservative", mit der üblichen Übersetzung "konservativ" aber nicht treffend bezeichnet) ist immer tätig, solange zu bearbeitende Tätigkei-

ten anstehen; sie erledigt genau die Tätigkeiten, die durch die Bedienwünsche der Kunden bezeichnet sind. Eine Station konstanter Bedienkapazität arbeitet, solange es zu bearbeitende Tätigkeiten gibt, mit konstanter Bearbeitungsgeschwindigkeit (Dimension: Arbeitseinheiten/Zeiteinheit).

konstante Bedienkapazität

Testfrage 2.2.9: Betrachten Sie Stationen mit den Bediendisziplinen

Testfrage

- FCFS (First Come First Served);
- HOL (Head Of the Line);
- RR (Round Robin).

Inwiefern, bzw. unter welchen Bedingungen, treffen die Gleichungen (2.2.7) für diese Stationen zu?

Wie man sich ebenfalls leicht überlegen kann, bleibt die Gleichheit $\overline{S} = \overline{B}$ auch erhalten, wenn das Beobachtungsintervall nicht (wie in Abb. 4.2.6) nur ein, sondern mehrere Belegt-Intervalle (voll) umfaßt. Sie geht allerdings verloren, wenn zu Beginn oder Ende des Beobachtungsintervalls die Station nicht leer ist. Mit "genügend großem" T wird die Abweichung dennoch klein, so daß immer noch $\overline{S} \approx \overline{B}$.

erweiterte Gültigkeit von (2.2.7f)

Um Ihnen zu zeigen, daß es durchaus Fälle gibt, in denen $\overline{S} \neq \overline{B}$, betrachten wir einen anderen (aber voll im Rahmen der Bedienstation nach Abb. 2.1.4 liegenden) Stationstyp (vgl. Abb. 2.2.10). Der Bediener bestehe aus beliebig vielen (untereinander identischen) Sub-Bedienern; ankommende Kunden finden jedenfalls unmittelbar einen freien Sub-Bediener, der ihre (jeweils alleinige, völlige) Bedienung übernimmt; der Warteraum ist demzufolge immer leer. Die Station sei arbeitserhaltend. Sie ist aber offensichtlich nicht von konstanter Bedienkapazität: Bei einer angenommenen Arbeitsgeschwindigkeit von r je Sub-Bediener ist vielmehr die Bedienkapazität der Gesamtstation bei Anwesenheit von einem Kunden gleich r, bei zwei anwesenden Kunden 2r (es werden ja beide gleichzeitig mit Geschwindigkeit r bedient), allgemein bei n>0 anwesenden Kunden n r. Hier ist sicher die für eine Station konstanter Bediengeschwindigkeit gültige Schlüsselbeziehung (2.2.7e) nicht gegeben: $\overline{B}$ hat zu $\overline{S}$ keinen einfachen Bezug. Wir werden diesen Stationstyp später wieder aufnehmen; für hier soll lediglich noch darauf hingewiesen werden, daß die Station (neben anderen Benennungen) deshalb den Namen "Verzögerungsstation" trägt, weil sie jeden Kunden genau nach Ablauf seiner Bedienzeit S_i entläßt, S_i also als reine Verzögerung im Ablauf des verursachenden Prozesses gesehen werden kann.

Verzögerungsstation

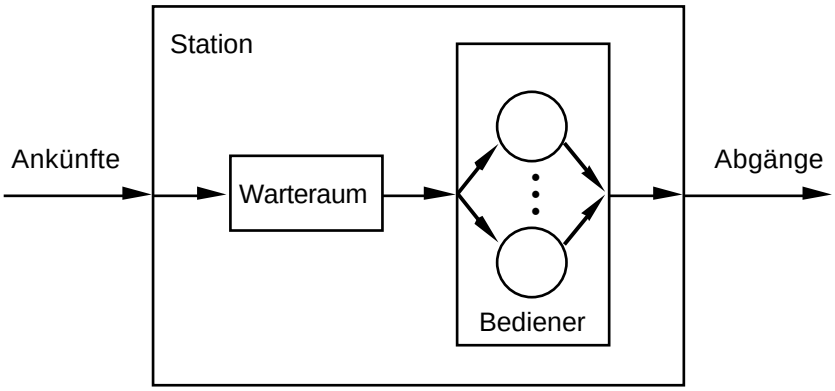


Abbildung 2.2.10: Verzögerungsstation

Testfrage

Testfrage 2.2.11: Nehmen Sie sich Beispiel 2.1.3 nochmals vor. Können Sie unter Einsatz einer Verzögerungsstation zu einer einfacheren Maschinen- und Last-(Modell-)Vorstellung kommen als dort verwendet? Hinweis: "Verschmelzen" Sie die Endgeräte.

Auslastung

Kehren wir nach diesen längeren Diskussionen zur Hauptlinie im Anschluß an Def. 2.2.3 samt zugehöriger Benennungen zurück. Sei zunächst erwähnt, daß für b auch die Benennung "Auslastung" gebräuchlich ist, was in der Interpretation "relativer Zeitanteil der Benutzung während des Beobachtungsintervalls" eine immer korrekte Benennung ist (ergibt sich direkt aus der Definition), in der für "Auslastung" ebenfalls gebräuchlichen Interpretation "relative Ausnutzung der Arbeitskapazität der Station" aber wieder nur für arbeitserhaltende Stationen konstanter Bedienkapazität zutrifft.

Aus Def. 2.2.3 erhält man unmittelbar

$$b = \frac{B}{T} = \frac{B}{C}\,\frac{C}{T} = \bar{B}\,c$$

Auslastungs-gesetz

kurz das sog. Auslastungsgesetz

(2.2.12) $b = c\ \bar{B}$

Betriebsgesetz

(2.2.12) ist unser erstes Beispiel für ein sog. Betriebsgesetz (operational law): Ein solches legt für jede mögliche Messung einen exakten Zusammenhang zwischen abgeleiteten Größen fest. Die Interpretation

Auslastung = Abgangsrate mittlere Bedienzeit

die ja nach kurzem Nachdenken sehr einleuchtend ist, ist mit der jetzt schon gewohnten Vorsicht zu genießen (als mittlere Bedienzeit kann in (2.2.12) nur in den bekannten Fällen $\bar{S}$ substituiert werden).

Betriebs-prinzipien

Neben meßbaren Basisgrößen, daraus abgeleiteten Größen und verschiedenen, diese Größen verbindenden Betriebsgesetzen (wir werden weitere kennenlernen) verwendet die Betriebsanalyse letztlich sog. Betriebsprinzipien (operational principles): Diese stellen zusätzliche Hypothesen (Annahmen) über Zusammenhänge zwischen den verwendeten Größen dar, die in manchen Fällen exakt, in vielen Fällen in guter Approximation gelten, immer aber aus den Messungen überprüfbar sind.

Unser erstes Beispiel für ein Betriebsprinzip beruht auf der plausiblen Annahme, daß für großes T (große Beobachtungs-Intervalle) die Zahl der Kundenankünfte an einer Station und die der Kundenabgänge von einer Station sich, relativ gesehen, wenig voneinander unterscheiden werden

Verkehrsfluß-gleichgewicht

(2.2.13a) $A \approx C$ Prinzip des
(2.2.13b) $a \approx c$ Verkehrsflußgleichgewichts

da ja (A-C)/C mit wachsendem T kleiner werden sollte. Unter Benutzung von (2.2.13) erhalten wir sofort eine zweite (approximative) Form unseres Auslastungsgesetzes (2.2.12)

(2.2.14) $b \approx a\ \bar{B}$

Wir werden im folgenden von der exakten Erfüllung der Betriebsprinzipien aus-

gehen und daher die "≈"-Zeichen nicht mehr explizit verwenden (obwohl dies eigentlich korrekt wäre und die substituierten "="-Zeichen oft nur approximativ - aber nachprüfbar!- gelten). Das als gültig vorausgesetzte Prinzip des Verkehrsflußgleichgewichts postuliert Gleichheit der Ankunftsrate a und der Abgangsrate c; als Benennung (für beide, als gleichwertig angenommenen Größen) hat sich die Bezeichnung "Durchsatz" eingebürgert.

Durchsatz

Verfeinern wir jetzt unsere Beobachtungen des Stationszustandes. Anstatt die gesamte Belegtzeit B zu messen, also die Summe der Zeiten mit mindestens einem anwesenden Kunden, beobachten wir die Gesamtdauer der Stationszustände bezüglich der Zahl anwesender Kunden, also Größen B(n), n=0,1,2,...

Definition 2.2.15: Basisgrößen Belegtzeiten

Belegtzeiten

B(n)	Summe aller Zeitintervalle (während T),
n=0,1,2,...	in denen die Station genau n Kunden beherbergte
	(vgl. auch Anmerkung im Anschluß an Def. 2.2.2).

Aus den Belegtzeiten ergeben sich Belegungsgrade (mit n Kunden) gemäß

Belegungsgrade

(2.2.16) $b(n) := B(n)/T \qquad n=0,1,2,..$

Offensichtlich ist

(2.2.17a) $B = \sum_{n>0} B(n)$

(2.2.17b) $T = \sum_{n\geq 0} B(n)$

Eine auf den Belegtzeiten basierende abgeleitete Größe ist

(2.2.18) $J := \sum_{n>0} n\, B(n) \qquad \left(= \sum_{n>0} n\, B(n) \right)$

welche die Summe aller (ins Beobachtungsintervall fallenden) Verweilzeiten aller im Beobachtungsintervall anwesenden Kunden mißt, und die wir "kumulative Verweilzeit" taufen wollen. Wir können uns die Größe J auch anschaulich machen anhand eines möglichen Geschehnisablaufs an der betrachteten Station (s. Abb. 2.2.19).

kumulative Verweilzeit

J entspricht genau der (schraffierten) Fläche unter der Treppenkurve. Die mittlere Höhe der Treppenkurve erhält man als

(2.2.20) $n = J/T$

Für n ist die Benennung "mittlere Kundenzahl" an der Station angemessen, wenn man bedenkt, daß man ungefähr denselben Wert erhält, wenn man die Kundenzahl während T hochfrequent äquidistant mißt und das Mittel bildet.

mittlere Kundenzahl

Wir können andererseits die kumulative Verweilzeit J auf alle beteiligten Kunden aufteilen und so die mittlere Verweilzeit pro Kunden ermitteln. Mitteln wir über die Zahl der Abgänge,

(2.2.21a) $\dot{J} = J/C$

dann ist diese "mittlere Verweilzeit" im Falle einer zu Beginn und zu Ende der Beobachtung leeren Station exakt gleich der "mittleren Kundenverweilzeit",

mittlere Verweilzeit

Belegungs-gebirge

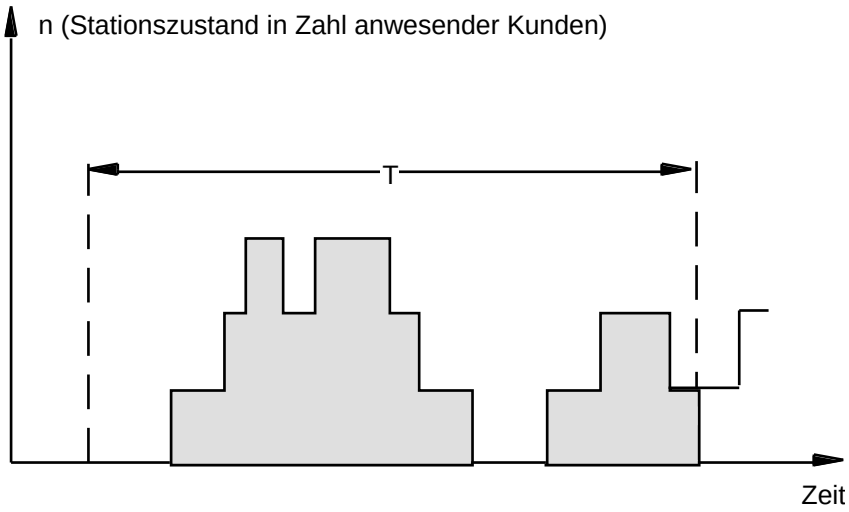


Abbildung 2.2.19: Belegungsgebirge

die man auf der Basis individueller Kundenverweilzeiten V_i aller beteiligten Kunden zu

(2.2.21b) $$\bar{V} = \frac{1}{C} \sum_{i=1}^{C} V_i$$

errechnen würde. Ist die Station zu Beginn oder Ende der Beobachtung nicht leer, dann ist $\bar{J}$ nur approximativ gleich $\bar{V}$, aber wieder mit der Tendenz, daß bei wachsendem T der relative Unterschied der beiden Größen sinkt. In jedem Fall erhalten wir aus (2.2.20, 2.2.21a, 2.2.3)

Little's Gesetz

(2.2.22a) $$n = \frac{J}{T} = \frac{J}{C} \cdot \frac{C}{T} = \dot{J} \cdot c$$ Little's Gesetz

ein wesentliches Betriebsgesetz, das eine Sonderform des Ihnen sicher wohlbekannten "Little'schen Theorems" darstellt und das bei Anwendung des Prinzips des Verkehrsflußgleichgewichts (2.2.13) - mit dadurch implizierter Annäherung von $\bar{J}$ und $\bar{V}$ - in die übliche Form

(2.2.22b) $$n = \bar{V} \cdot a$$

übergeht.

Gesamtnetz

Nach diesen Überlegungen bzgl. einer beliebigen Einzelstation des zu betrachtenden Verkehrsnetzes wenden wir uns jetzt dem Gesamtnetz zu. Ich erinnere an die weiterhin gültige Maschinenstruktur nach Abb. 2.1.1. Die Anzahl der Stationen sei M. Unsere (potentiell möglichen) Messungen während des Betriebs umfassen folgende Basisgrößen:

Basisgrößen

Definition 2.2.23: Basisgrößen

T Gesamtdauer der Beobachtung

sowie die während des Beobachtungsintervalls an den Stationen gemessenen

- A_j Zahl der Kundenankünfte an Station j; j=1,2,…,M
- C_j Zahl der Kundenabgänge von Station j; j=1,2,…,M
- B_j Belegtzeit der Station j (Gesamtzeit mit mindestens einem Kunden); j=1,2,…,M

C_{ij} Zahl der Abgänge von Station i, die nach Station j gehen ("Übergangszahl"); i,j=1,2,…,M
C_{0j} Zahl der Kundeneintritte ins Netz, die nach Station j gehen; j=1,2,…,M
C_{i0} Zahl der Abgänge von Station i, die das Netz verlassen; i=1,2,…,M

Umgebung

Bei den Benennungen C_{0j}, C_{i0} haben wir die Umgebung als eine Art zusätzliche Station mit Index 0 behandelt; sie tritt natürlich nur bei offenem Netz in Erscheinung. Der Vollständigkeit halber können wir aber in diesem Fall mit C_0 noch die Gesamtzahl der "Abgänge aus der Umgebung" (d.h. der Ankünfte am Netz), mit A_0 die Zahl der "Ankünfte in der Umgebung" (d.h. der Abgänge vom Netz) bezeichnen und sinnvollerweise $C_{00}=0$ festsetzen (wir betrachten ja niemanden, der das Netz nicht berührt).

Analog zur Ableitung der Größen 2.2.3 und etwas darüber hinausgehend ergeben sich:

abgeleitete Größen

Definition 2.2.24: Abgeleitete Größen

$a_j = A_j/T$ Ankunftsrate Station j
$c_j = C_j/T$ Abgangsrate Station j
$\underline{b}_j = B_j/T$ Belegungsgrad Station j
$\overline{B}_j := B_j/C_j$ Mittlere Bedienzeit Station j
$c_{ij} = C_{ij}/T$ Übergangsrate Station i - Station j
$h_{ij} = C_{ij}/C_i$ Relative Übergangshäufigkeit Station i -> Station j (relativer Anteil der Übergänge i-j bezüglich der Gesamtabgänge i)

wobei die Benennungen mit der bei der Einzelstation eingeübten Sorgfalt zu verwenden sind.

Im Verhältnis der Stationen zueinander bemerken wir zunächst, daß jede Kundenbewegung einerseits als Abgang von einer Station, andererseits als Zugang zu einer Station gezählt wird, so daß

(2.2.25) $$A_j = \sum_{i=0}^{M} C_{ij} \qquad j=0,1,\ldots,M$$

Setzen wir nun Verkehrsflußgleichgewicht gemäß (2.2.13) voraus, also

(2.2.26) $$A_i = C_i \qquad i=0,1,\ldots,M$$

dann wird aus (2.2.25)

$$C_j = \sum_{i=0}^{M} C_{ij} = \sum_{i=0}^{M} \frac{C_{ij}}{C_i}\ C_i$$

und mit Def. 2.2.24:

(2.2.27a) $$C_j = \sum_{i=0}^{M} h_{ij}\ C_i \qquad j=0,1,\ldots,M$$

bzw.

(2.2.27b) $$c_j = \sum_{i=0}^{M} h_{ij}\ c_i \qquad j=0,1,..,M$$

Verkehrsfluß-gleichgewicht

(2.2.27) trägt den Namen "Formel des Verkehrsflußgleichgewichts" im Netz. Die Formel drückt, im Sinne ihrer Herleitung, eine Beziehung zwischen den beobachteten bzw. aus Beobachtungen abgeleiteten Größen c., h. aus, die mit wachsendem Beobachtungsintervall immer genauer eingehalten wird. Im Sinne zwingender wechselseitiger Abhängigkeiten zwischen den Größen c., h. interpretiert, führt sie zu interessanten, praktisch relevanten Folgerungen.

Nehmen wir einmal an, wir wüßten über die relativen Übergangshäufigkeiten h. Bescheid, ohne sie aus den Abgangszahlen gemäß Def. 2.2.23/2.2.24 bestimmen zu müssen. Dies ist eine durchaus realistische Annahme, liegen doch den Übergangshäufigkeiten die Bearbeitungswünsche der treibenden Lastprozesse zugrunde, die wir in Abschnitt 2.1 durch geeignete Prozeßmuster charakterisiert hatten. Konkreter gesprochen, sollten sich die h. direkt aus den dort ausgedrückten Reihenfolgevorschriften ergeben; so etwa

in Bsp. 2.1.3 zu $h_{terminal\_i,rechner} = 1$
oder in Bsp. 2.1.4 zu $h_{leitwerk,rechenwerk} = r$.

Eine naheliegende Frage ist, ob nun (mit dieser Kenntnis der h.) die Abgangszahlen C. aus Gl. (2.2.27) direkt, also ohne Messungen bestimmbar sind; bzw. besser noch die Abgangsraten c., da sie von der Länge T des Beobachtungsintervalls unabhängig sind.

Zusammenhang der Abgangsraten

Das Gleichungssystem (2.2.27b) hat, leicht umgeschrieben, die Form

(2.2.27c) $$\sum_{i=0,\neq j}^{M} h_{ij}\ c_i + (h_{jj}-1)\ c_j = 0 \qquad j=0,1,\ldots,M$$

oder in Vektor-/Matrix-Schreibweise, mit $\underline{c} = (c_0,c_1,...,c_M)^T$ als Vektor der Abgangsraten, $H = (h_{ij};\ i,j=0,1,\ldots,M)$ als Matrix der relativen Übergangshäufigkeiten, $\underline{0}$ als Nullvektor und E als Einheitsmatrix passender Dimension

$$\underline{c}^T(H-E) = \underline{0}^T$$

bzw. letztlich, mit der Abkürzung $H^* = (h^*_{ij}) = H-E$

(2.2.27d) $$c^T H^* = 0^T$$

Dies ist ein homogenes lineares Gleichungssystem, das nur dann eine nichttriviale Lösung besitzt, wenn die Determinante von H* verschwindet - also det(H*) = 0 - oder, anders ausgedrückt, der Rang von H* kleiner ist als die Dimension von H* - also rang(H*) < dim(H*) = M+1. Diese Bedingung ist tatsächlich immer erfüllt, da nach Def. 2.2.24

$$\begin{aligned}\sum_{j=0}^{M} h^*_{ij} &= \sum_{j=0}^{M} h_{ij} - 1 \qquad i=0,1,\ldots,M\\ &= \sum_{j=0}^{M} C_{ij}/C_i - 1\\ &= C_i/C_i - 1\\ &= 0 \qquad\qquad i=0,1,\ldots,M\end{aligned}$$

alle Zeilensummen von H* identisch verschwinden. Gehen wir der Lösbarkeit von (2.2.27d) weiter nach: Die Lösung eines homogenen linearen Gleichungssystems mit Koeffizientenmatrix H* und Dimension M+1 besitzt

$$s = (M+1) - rang(H^*)$$

Freiheitsgrade, d.h. s der Unbekannten (hier der c.) sind frei wählbar, die restlichen M+1-s sind aus diesen s errechenbar. Man kann sich leicht überlegen, daß i. allg. rang(H*)=M: Wäre dem nicht so, dann müßte jede Unterdeterminante der Dimension M aus Matrix H* verschwinden. Nehmen wir beispielsweise die ersten M Spalten von H*, bedeutete dies, daß eine Linearform L existieren müßte derart daß

$$L(h^*_{ij};j=0,1,\ldots,M-1) = 0 \qquad i=0,1,\ldots,M$$

Dies ist aber nicht generell möglich: In L waren die Elemente h^*_{iM} nicht berücksichtigt, d.h. ohne Belang. Diese Elemente sind aber "frei setzbar" in dem Sinne, daß sie für jedes Problem einen anderen Wert haben können (sehen Sie die Definition der h*.. bzw. h.. an) - dies für jede Zeile i je unabhängig. Die Setzung der h^*_{iM} führt über den erkannten Zusammenhang $\Sigma h^*_{ij} = 0$ zu (dem Problem entsprechenden) Änderungen an den $h^*_{ij};j=0,1,\ldots,M-1$; je zeilenweise unabhängig läßt sich damit obige Forderung an die Linearform L verletzen - es kann ein solches L nicht geben. rang(H*) kann demnach höchstens für einzelne, speziell gelagerte Probleme kleiner als M sein, nicht aber im allgemeinen.

Fallunterscheidung

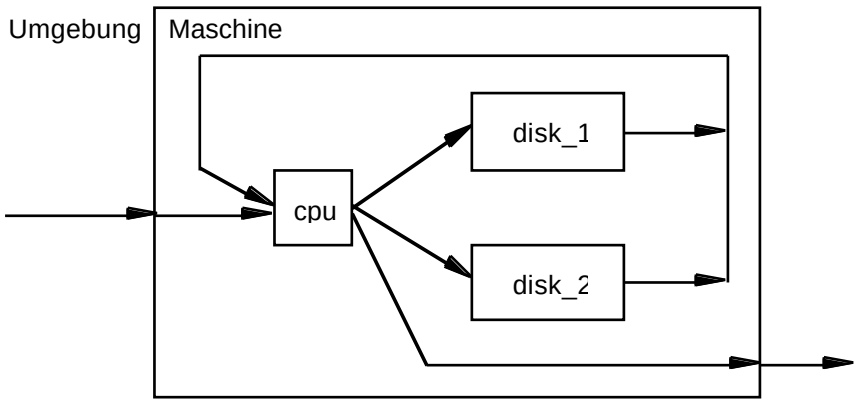
Treffen wir nach diesen Überlegungen eine Fallunterscheidung zwischen offenen und geschlossenen Systemen:

offenes System

(i) Offenes System:
Hier gibt es laut Definition Übergänge zwischen Umgebung und System und meßbare, nicht verschwindende Systemankunfts- bzw. Systemabgangsraten c_0 (Verkehrsflußgleichgewicht!). (2.2.27c) hat (i.allg.) genau einen Freiheitsgrad. Was liegt näher, als c_0 (die von der Umgebung erzwungene, treibende Zugangsrate) als Systemparameter zu betrachten! Damit liegen aber gemäß (2.2.27c) alle anderen Durchsätze (Abgangsrate = Ankunftsrate = "Durchsatz") c_i, i=1,2,…,M, als Funktionen von c_0 fest: Aus Kenntnis von c_0 lassen sich alle Verkehrsflüsse im Netz errechnen. Dies setzte allerdings Kenntnis der relativen Übergangshäufigkeiten h. voraus und implizit ein bisher unerwähntes Betriebsprinzip, das der "Homogenität des Verkehrsflusses": Die relativen Übergangshäufigkeiten h. werden als konstant angesehen, insbesondere unbeeinflußt von der absoluten Verkehrsstärke im Netz. Solange aber, wie angenommen, die h. unmittelbare Folge zwingender Prozeßmuster sind (nach denen sich die Maschine gefälligst zu richten hat!), ist daran sicher nichts auszusetzen.

Testfrage

Testfrage 2.2.28: Ein idealisiertes Rechensystem bestehe (vgl. folgende Skizze) aus 3 Stationen, einer Zentraleinheit cpu und zwei Ein/Ausgabeeinheiten disk_1 und disk_2. Die Maschinenstruktur habe inklusive aller Übergangsmöglichkeiten obiges Aussehen. Die Maschine werde von einem Strom von Stapelaufträgen belastet, deren jeder eine Folge von Anforderungen der Form cpu-disk_i-cpu-disk_j-... stellt. Jeder Auftrag berührt cpu genau 31-mal. Die E/A-Aufträge gelten in zwei Drittel aller Fälle der Einheit disk_1, die restlichen der disk_2. Ermitteln Sie die Durchsätze an cpu, disk_1, disk_2 in Abhängigkeit von der Ankunftsrate der Stapelaufträge.



geschlossenes System

(ii) Geschlossenes System:
Hier gibt es laut Definition keine Übergänge Umgebung/System, d.h. $c_0=a_0=0$. (2.2.27c) erhält, entsprechend reduziert, die Form

$$\sum_{i=1,\neq j}^{M} h_{ij}\, c_i \;+\; (h_{jj}-1)\, c_j \;=\; 0 \qquad\qquad j=1,2,\ldots,M$$

Analog den Überlegungen zu (2.2.27d) ist dies ein homogenes lineares Gleichungssystem der Dimension M mit Rang M-1. Zur Ausfüllung des verbleibenden Freiheitsgrades bietet sich kein "natürlicher" Parameter an (wie das im Fall des offenen Systems die Systemzugangsrate c_0 war). Somit ist auch keine eindeutige Lösung des Gleichungssystems verfügbar. Wie können wir die obigen Beziehungen dennoch nutzen?

Betrachten wir offene und geschlossene Systeme wieder gemeinsam, also Gleichungssysteme (2.2.27). Greifen wir willkürlich eine Station j* heraus und definieren die Größen

(2.2.29a) $$X_j = \frac{C_j}{C_{j^*}} \qquad\qquad j=0,1,\ldots,M$$

Wegen der Definitionen 2.2.24 und wegen des angenommenen Verkehrsflußgleichgewichts gilt offenbar auch

(2.2.29b) $$X_j = \frac{A_j}{A_{j^*}} = \frac{c_j}{c_{j^*}} = \frac{a_j}{a_{j^*}}$$

relative Besuchszahlen

Nach Def. (2.2.29a) ist X_j die "Zahl der Abgänge von Station j pro Abgang von Station j*" und trägt den Namen "relative Besuchszahl" (relative visit count). Substituiert man in (2.2.27c) die c_j entsprechend (2.2.29b), also gemäß

(2.2.30) $$c_j = X_j \;\; c_{j^*} \qquad\qquad j=0,1,\ldots,M$$

dann ergibt sich

$$\sum_{i=0,\neq j}^{M} h_{ij}\, X_i\, c_{j^*} \;+\; (h_{jj}-1)\, X_j\, c_{j^*} \;=\; 0 \qquad j=0,1,\ldots,M$$

bzw. ($c_{j^*}\neq 0$ vorausgesetzt)

(2.2.31a) $$\sum_{i=0,\neq j}^{M} h_{ij}\, X_i \;+\; (h_{jj}-1)\, X_j \;=\; 0 \qquad\qquad j=0,1,\ldots,M$$

wozu noch das offensichtlich richtige

(2.2.31b) $$X_{j^*} = \frac{C_{j^*}}{C_{j^*}} = 1$$

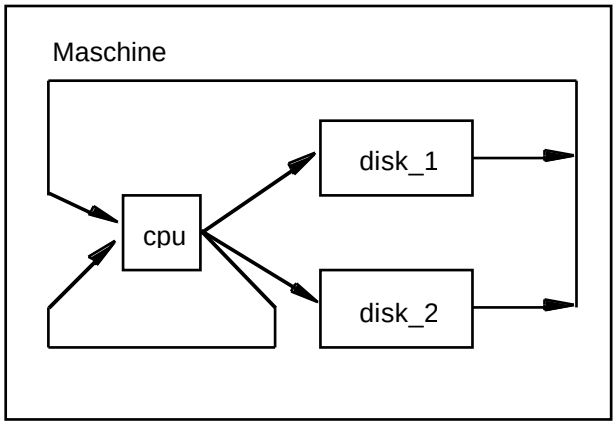
tritt. (2.2.31a) ist mit (2.2.27c) identisch bis auf die zu errechnenden Unbekannten (dort Durchsätze c., hier relative Besuchszahlen X.). Auf (2.2.31a) treffen alle bzgl. (2.2.27c) getroffenen Bemerkungen zu. Insbesondere läßt sich wegen der linearen Abhängigkeit der Gleichungen (2.2.31a) eine beliebige davon streichen - und durch (2.2.31b) ersetzen, wodurch (2.2.31) insgesamt ein nicht-homogenes lineares Gleichungssystem darstellt, das i.allg. (bis auf rein numerisch bedingte Sonderfälle) eine eindeutige Lösung besitzt.

Bestimmung der relativen Besuchszahlen

Wir können also bei Kenntnis der Matrix H der Übergangshäufigkeiten alle relativen Besuchszahlen X. (bezüglich einer beliebigen Station j*) explizit bestimmen. In einer meßtechnischen Anwendung der Beziehungen bedeutet dies, daß die Messung der Abgangsrate an einer beliebigen Station j* genügt, um alle Durchsätze mittels (2.2.30) zu erhalten (die X_j konnten wir ja errechnen!).

Testfrage

Testfrage 2.2.32: Betrachten Sie erneut Testfrage 2.2.28. In Abänderung der Aufgabe nehmen wir an, daß das System (vom Betriebssystem her) über einen vorgegebenen maximalen Multiprogrammierungsgrad MP in der Kundenaufnahme beschränkt ist und "unter Vollast" betrieben werde. Der Modellierungs-"Trick", der hierfür üblicherweise eingesetzt wird, besteht in folgender Abänderung der Maschine



und der Annahme, daß ein "abgearbeiteter" Stapelauftrag (der eigentlich das System verläßt) wegen bestehender Vollast sofort durch einen neuen gleicher Art ersetzt wird (im Modell also "wieder von vorn anfängt", im speziellen Beispiel von cpu-Ausgang nach cpu-Eingang geht). Alle sonstigen Annahmen von Testfrage 2.2.28 bleiben bestehen. Sie messen 40 Abgänge/ZE an disk_1. Wie sind die Durchsätze aller Stationen? Wieviel Stapelaufträge werden je ZE abgewickelt? Sei bei dieser Messung MP=5 festgehalten. Wielange dauert im Mittel die gesamte Bearbeitung eines Stapelauftrags (denken Sie an Little!)?

Belastbarkeit

Das Konzept der relativen Besuchszahlen führt in seiner praktisch wohl wesentlichsten Anwendung zu Erkenntnissen über die maximale Belastbarkeit des Netzes und über jene Stationen, welche die Belastbarkeit begrenzen. Kehren wir zum Auslastungsgesetz (2.2.12) zurück, das ja für jede Station des Netzes gilt, so

daß also

$$b_j = \bar{B}_j c_j \qquad j=1,2,\ldots,M$$

Gemäß (2.2.30) ist damit auch

$$b_j = \bar{B}_j X_j c_j^* \qquad j=1,2,\ldots,M$$

bzw.

(2.2.33a) $$\frac{b_j}{\bar{B}_j X_j} = c_j^* \qquad j=1,2,\ldots,M$$

Das bedeutet zum einen, daß der linksstehende Ausdruck systemweit invariant ist:

systemweite Invarianz

(2.2.33b) $$\frac{b_1}{\bar{B}_1 X_1} = \frac{b_2}{\bar{B}_2 X_2} = \ldots = \frac{b_M}{\bar{B}_M X_M}$$

Berücksichtigen wir weiter, daß laut Definition 2.2.3 kein b_j den Wert 1 übersteigen kann, dann bedeutet (2.2.33a) zum anderen, daß

(2.2.33c) $$c_j^* \le \frac{1}{\bar{B}_j X_j} \qquad j=1,2,\ldots,M$$

das heißt

physikalische Durchsatzgrenze

$$c_j^* \le \min_{j=1,2,\ldots,M} \frac{1}{\bar{B}_j X_j}$$

gelten muß im Sinne einer physikalischen Grenze des Durchsatzes.

Voraussetzungen für Flaschenhalsbetrachtungen

Für die folgenden Überlegungen setzen wir voraus, daß alle Stationen des Netzes arbeitserhaltend und von konstanter Bedienkapazität r=1 im Sinne der Def. 2.2.8 sind. Damit dürfen wir für alle Stationen

$$\bar{B}_i = \bar{S}_i \qquad i=1,2,\ldots,M$$

setzen, da ja (wie vorne diskutiert) die (aus Messungen ermittelten) Größen $\bar{B}_i$ mit den (auf den Bedienwünschen $\bar{W}_i$ basierenden) mittleren Bedienzeiten $\bar{S}_i$ übereinstimmen. In ähnlicher Weise wie anläßlich der Entwicklung der Gleichungen (2.2.27) für die Übergangshäufigkeiten h. diskutiert, sind jetzt auch die mittleren Bedienzeiten $\bar{S}_i$ in vielen Fällen bereits aus der Beschreibung der treibenden Lastprozesse bekannt, u.U. in Prozeßmustern genau festgelegt. Selbst wenn die Größe der $\bar{S}_i$ nicht bekannt ist, dürfen wir doch davon ausgehen, daß sie allein auf den Kundenwünschen beruhen, daher von Gegebenheiten des Betriebs unbeeinflußt sind und von einer Messung zur anderen (bei u.U. unterschiedlichen Verkehrsstärken im Netz) unverändert bleiben.

Trennen wir nach diesen Vorbemerkungen wieder unsere Diskussion offener von der geschlossener Systeme:

offenes System

(i) Offenes System:
Wir hatten erkannt, daß der gesamte Verkehr im Netz durch den (von außen erzwungenen) Zugangsstrom der Rate c_0 bestimmt wird. Wir tragen dieser Sonderrolle der Umgebung dadurch Rechnung, daß wir für alle Gleichungen ab (2.2.29) j*=0 wählen. Insbesondere erkennen wir aus (2.2.30), daß alle Stationsdurchsätze gemäß

$$c_i = X_i \cdot c_0 \qquad i=1(1)M$$

linear mit c_0 wachsen (die X_i sind auf j*=0 bezogen und damit X_0=1), solange dies physikalisch möglich ist. Wir entnehmen jetzt der Gl. (2.2.33c), daß

$$c_0 \leq \min_{i=1,2,\ldots,M} \frac{1}{\bar{S}_i X_i}$$

wobei wir $\bar{S}$. für $\bar{B}$. substituiert haben wegen der Annahme der Identität der beiden Größen sowie der des fehlenden Einflusses der Verkehrsstärke auf die $\bar{S}$. Die physikalisch maximale Zugangsrate c_0 (der maximale Systemdurchsatz) ist damit aus den mittleren Bedienzeiten $\bar{S}$. und den relativen Besuchszahlen X. direkt bestimmbar (wobei die Bestimmung der X. ihrerseits die Kenntnis der Übergangshäufigkeiten h. voraussetzte).

maximaler Systemdurchsatz

Testfrage 2.2.34: Was geschieht, wenn die Zugangsrate c_0 über den ermittelten Maximalwert hinaus gesteigert wird?

Testfrage

Führen wir uns die Abhängigkeit der Stationsdurchsätze c_i, i=1,2,…,M, von der Systemzugangsrate c_0 der Deutlichkeit halber graphisch vor Augen, wobei als Beispiel M=3 gewählt ist und $\bar{S}_1X_1 > \bar{S}_2X_2 > \bar{S}_3X_3$, sowie $X_2 > X_1$, $X_3 < X_1$ gelte (s. Abb. 2.2.35). Alle Stationsdurchsätze steigen linear mit c_0, alle besitzen (entsprechend der Begrenzung von c_0) einen spezifischen, maximal erreichbaren Wert.

Stationsdurchsätze

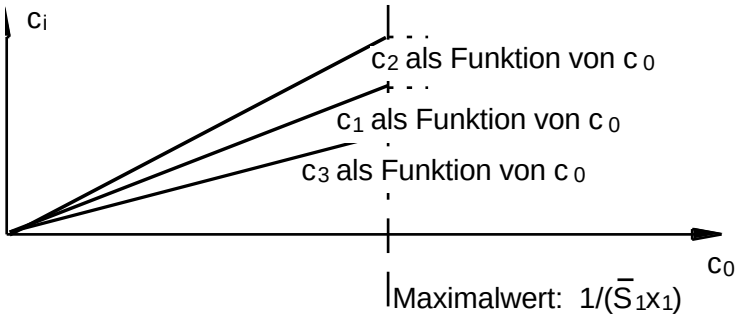


Abbildung 2.2.35: Durchsätze als Funktionen der Zugangsrate

Die Darstellung wird noch aufschlußreicher, wenn wir statt der Durchsätze c. die Auslastungen b. in ihrer Abhängigkeit von c_0 betrachten. Laut (2.2.33a) gilt ja

Stationsauslastungen

$$b_i = \bar{S}_i X_i c_0 \quad i=1,2,\ldots,M$$

was wegen der Beschränkung von c_0 durch die Station mit maximalem $\bar{S}$.X. die Auslastungen aller anderen Stationen auf Werte

$$b_i \leq \bar{S}_i X_i c_{0,max} = \frac{\bar{S}_i X_i}{(\bar{S}. X.)_{max}}$$
$$\leq 1$$

beschränkt. Graphisch ergibt sich mit obigen Beispielsannahmen Abb. 2.2.36. Wir interpretieren: Alle Stationsauslastungen b. steigen linear mit dem Systemdurchsatz c_0 bis zu einem Maximalwert. Für (i.allg. genau:) eine Station k wird b_k=1 erreicht. Diese Station (sie weist das maximale $\bar{S}$ X aller Stationen auf) be-

Interpretation

Flaschenhals

grenzt eine weitere Erhöhung von c_0: Sie wird daher als "Flaschenhals" (englisch: "bottleneck") des Systems bezeichnet. Alle anderen Stationen sind zwingend auf Auslastungswerte $b_i < 1$, i=1,2,…,M, i≠k, festgelegt. Diese Interpretation hat offensichtlich zentrale praktische Bedeutung, wie in folgender Testfrage überdeutlich werden sollte.

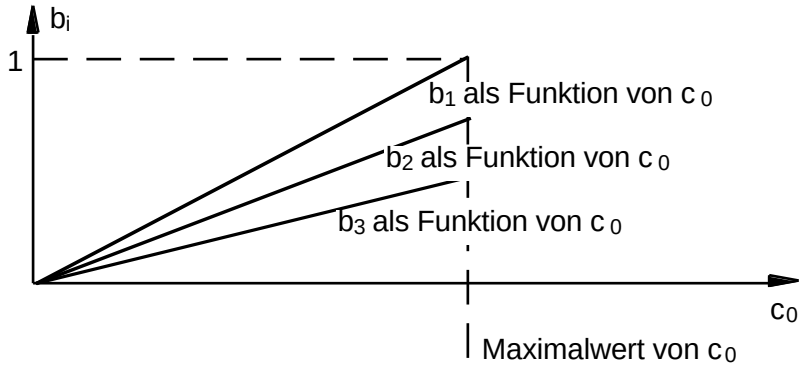


Abbildung 2.2.36: Auslastungen als Funktionen der Zugangsrate

Testfrage

Testfrage 2.2.37: Eine Leistungsverbesserung eines Rechensystems läßt sich in vielen Fällen durch Ersatz einer Maschinenkomponente durch eine funktional gleichwertige, aber schnellere, andere erreichen. In unserer Modellwelt heißt dies, daß die Arbeitskapazität r_i einer Station i vergrößert wird, so daß (bei gleichbleibendem mittlerem Bedienwunsch $\overline{W}_i$) die mittlere Bedienzeit $\overline{S}_i$ sinkt. Können Sie durch eine derartige Maßnahme immer den maximalen Systemdurchsatz c_0 erhöhen?

geschlossenes System

(ii) Geschlossenes System:
Wir hatten schon diskutiert, daß sich bei geschlossenen Systemen unter den betrachteten Größen keine als natürlicher Parameter anbietet, um den verbleibenden Freiheitsgrad des Gleichungssystems (2.2.27) auszufüllen und es einer Lösung zuzuführen. Andererseits wissen wir (machen uns dies aber erst jetzt richtig bewußt), daß in einem geschlossenen System aufgrund der fehlenden Zu- und Abgänge eine konstante Anzahl von Kunden zirkuliert. Sei N die Zahl dieser Kunden. So wie beim offenen System der als unabhängige Größe angenommene Systemdurchsatz c_0 die Verkehrsdichte im Netz bestimmte, sollte man annehmen, daß beim geschlossenen System die Zahl N insgesamt anwesender Kunden (die "Gesamtpopulation") ihren Einfluß auf die Verkehrsdichte im Netz hat, sie evt. sogar völlig bestimmt. Fassen wir also im geschlossenen System die Gesamtpopulation N als zentralen Parameter auf und überlegen uns ihren Einfluß auf das Verkehrsgeschehen. Dazu folgender

Gesamtpopulation

Satz 2.2.38: Sei mit $b_i(N)$ die Auslastung der Station i, i=1,2,…,M, eines geschlossenen Systems mit einer Gesamtpopulation N, N=0,1,2,..., bezeichnet (und mögen die für den jetzigen Kontext getroffenen Annahmen zutreffen!). Dann gilt

$$b_i(N) \le b_i(N+1) \qquad \begin{matrix} i=1,2,\ldots,M; \\ N=0,1,2,... \end{matrix}$$

und für mindestens eine Station k

$$b_k(N) < b_k(N+1)$$

In Worten: Bei Einbringen eines zusätzlichen Kunden in das (ansonsten unveränderte) System wird die Auslastung keiner Station (im Vergleich zu den vor Einbringen beobachteten Werten) sinken, für mindestens eine Station steigen.

Der Satz entspricht sicher der Intuition. Der interessierte Leser findet einen Beweis (allerdings in stochastischem Kontext) z.B. in ChLa74. Der Beweis ist länglich und in seiner Beweisidee hier uninteressant, so daß wir den Satz ohne Beweis übernehmen.

Erinnern wir uns, daß nach (2.2.33b) die Größen $b_j/(\bar{B}_j X_j)$ systemweit invariant waren, z.B. den Wert K hatten. Dabei hatte (unausgesprochen) die Gesamtpopulation irgendeinen festen Wert. Bei Variation der Population N muß sich, wenn wir laut Satz 2.2.38 von durch diese Variation veränderten Durchsätzen $b_j(N)$ ausgehen (und bei entsprechend getroffener Annahmen unveränderlicher $\bar{B}.,X.$), die Systeminvariante K ebenfalls ändern, so daß aus (2.2.33b) die Beziehungen

(2.2.39) $$\frac{b_j(N)}{\bar{B}_j X_j} = K(N) \qquad j=1,2,\ldots,M;\ N=0,1,2,\ldots$$

folgern. Da nach wie vor (siehe Diskussion des offenen Systems) keine Stationsauslastung den Wert 1 übersteigen kann, ist K(N) in seiner Größe beschränkt. Wenn wir entsprechend Satz 2.2.38 und (2.2.33b) von monoton mit N steigenden Auslastungen ausgehen, sollte auch K(N) gemäß (2.2.39) monoton mit N steigen und sein Maximum bei über alle Grenzen wachsendem N erreichen. Schreiben wir kurz K(∞) für lim(K(N);N->∞), dann erhalten wir zusammengefaßt

(2.2.40) $$K(\infty) \le \min_{j=1(1)M} \frac{1}{\bar{B}_j X_j}$$

und daraus mit (2.2.39) und der Abkürzung

$$b_i(\infty) = \lim(b_i(N);N->\infty)$$

die Beziehung:

(2.2.41) $$b_i(\infty) = \bar{B}_i X_i K(\infty) \qquad i = 1,2,\ldots,M$$
$$\le \bar{B}_i X_i \min_{j=1(1)M} \frac{1}{\bar{B}_j X_j}$$

Unter der weiteren Annahme, daß für "unendlich großes" N die Grenzwerte tatsächlich erreicht werden, sollten wir die Skizze der Abb. 2.2.42 als prinzipielle *Auslastungen*

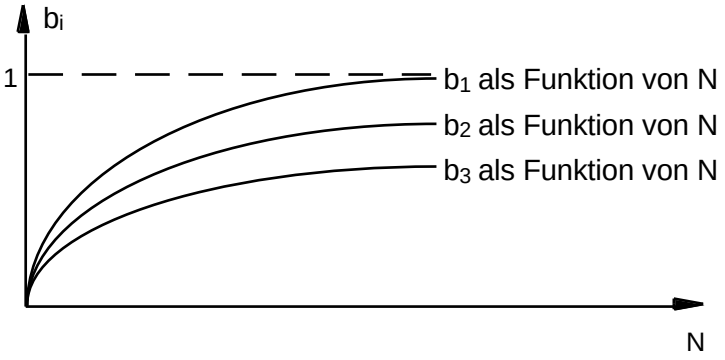


Abbildung 2.2.42: Auslastungen als Funktionen der Population

Abhängigkeit der Stationsauslastungen von der Gesamtpopulation festhalten, wobei (wie für Abb. 2.2.35/2.2.36) M=3, $\bar{S}_1X_1 > \bar{S}_2X_2 > \bar{S}_3X_3$ gelte.

Interpretation

Diese Skizze ist als analog der Abb. 2.2.36 des offenen Systems zu sehen. Beide Abbildungen zeigen die Abhängigkeit der Stationsauslastungen vom zentralen, verkehrsbestimmenden Systemparameter (dort Systemdurchsatz c_0, hier Systempopulation N) auf, wobei allerdings die Funktionsverläufe der Abb. 2.2.42 nur als Prinzipskizzen zu verstehen sind (wir wissen nur, daß monotones Steigen mit endlicher Asymptote vorliegt). Auch die Interpretation der Skizze ist analog der von Abb. 2.2.36 zu sehen: Alle Stationsauslastungen steigen monoton mit der Systempopulation N gegen einen Grenzwert. Dieser liegt für (i.allg. genau:) eine Station k bei $b_k(\infty)=1$. Diese Station (sie weist das maximale $\bar{S}$. X. aller Stationen auf) begrenzt die Auslastungen aller Stationen im Netz und wirkt, wie beim offenen System, als Flaschenhals; insbesondere sind die Auslastungen der Stationen i ($i \neq k$) damit auf Werte $b_i<1$ festgelegt. Die zentrale praktische Bedeutung dieser Aussage ist offensichtlich.

Folgefragen

Wir haben mit diesen Ergebnissen ein beträchtliches Verständnis für die Vorgänge in einem (als Verkehrsnetz modellierten) Rechensystem gewonnen. Gleichzeitig eröffnen sich bisher unbeantwortete Folgefragen: Wie ist der genaue Verlauf der $b_i(N)$ im geschlossenen Netz? Und (ein Fragenkomplex, den wir bisher ganz ausgespart haben) wie bestimmen wir die Verweilzeiten (d.h. die Abwicklungszeiten für eine Anforderung) der Kunden an den Stationen? Ferner: Was ist, wenn es mehrere "Sorten" von Kunden (d.h. mehrere Prozeßmuster) gibt? Auch für diese Fragen stellt die Betriebsanalyse Antworten bereit, die über zusätzlich eingeführte Betriebsprinzipien gewonnen werden. Die wesentlichen weiterführenden Konzepte sind die des "Zustands" des dynamischen Systems Verkehrsnetz (grob gesprochen die Frage "wieviele Kunden sind an welcher Station"?) und die der Verfolgung der Zustandsänderungen. Diese Konzepte werden wir (allerdings auf anderer Modellierungsbasis) in den folgenden Abschnitten ohnehin kennenlernen, so daß wir unser Studium der Betriebsanalyse hier abbrechen (dem interessierten Leser sei empfohlen, sich nach Bearbeitung der folgenden "stochastischen" Modellierungskapitel die weiterführenden Teile der angegebenen Betriebsanalyse-Literatur DeBu78, Rood79 vorzunehmen; auch das Lehrbuch LZGS84 argumentiert weitgehend auf der Basis der operational analysis).

2.3 Stochastische Modelle

Rückblick

Wir hatten uns in Abschn. 2.1 eine qualitative Vorstellung von Verkehrsnetzen erarbeitet. Ziel dabei war, ein Gerüst für die Bildung von analytischen Leistungsmodellen bereitzustellen. Wir haben dieses Gerüst in Abschn. 2.2 interpretiert mittels der Technik der Betriebsanalyse. Dabei haben wir uns auf (reale und/oder hypothetische) Messungen während eines Betriebsablaufes abgestützt und Beziehungen abgeleitet, die zwischen verschiedenen Meßgrößen exakt oder in guter Näherung gelten.

neuer Ansatz

Im Unterschied zur Betriebsanalyse, die (ich wiederhole:) jeweils einen einzelnen Betriebsablauf betrachtete, gehen die stochastischen Modelle von vornherein von einer unbeschränkten Menge möglicher Betriebsabläufe aus. Damit lassen sich allerdings Messungen nicht mehr an den Anfang der Überlegungen stellen (wer kann schon unbeschränkt oft messen?); vielmehr wird eine Betrachtung der Regeln und Gesetze, denen Betriebsabläufe unterliegen, als Grundlage verwendet; eine Betrachtung der Regeln also, nach denen letztlich die Werte der interessierenden Beobachtungs- (Meß-) Größen entstehen ("erzeugt werden"). Für die Formulierung der Regeln spielt der Detaillierungsgrad der Untersuchungen natürlich eine wesentliche Rolle. Wir haben uns in dieser Beziehung auf der Ebene der Verkehrsnetze bewegt und wollen dies Detaillierungsniveau auch beibehalten - sind aber dadurch genötigt, die Mehrzahl der Regeln nicht "mit Sicherheit" (also deterministisch) zu formulieren, sondern "zufallsbedingt" variierend (und damit stochastisch).

Erinnerung an Verkehrsnetze

Welche Regeln und Gesetze müssen nun konkret formuliert werden, um die Betriebsabläufe in einem Verkehrsnetz festzulegen? Erinnern wir uns an Abschn. 2.1: Wir arbeiten mit einer Menge von Stationen, die durch Wege zu einem Netz verbunden sind; in diesem Netz aus Stationen bewegen sich Kunden entsprechend den Regeln ihrer Prozeßmuster. Die Stationen dienten der Darstellung von Betriebsmitteln, die sich bewegenden Kunden der Erfassung der in Bearbeitung befindlichen Lasteinheiten (Aufträgen,Tasks), die Prozeßmuster der Festlegung der Bearbeitungswünsche jedes zugehörigen Lastprozesses (nach Bedienungsort, Bedienungsumfang und Bedienungsreihenfolge). Wir haben dieses allgemeine Bild eingeschränkt: Wir ließen nur sehr simple Bedienstationen zu, an die Bedienwünsche in Form des für die Bedienung erforderlichen Zeitaufwandes gestellt werden konnten (bzw. in Form eines eindimensionalen Arbeitsaufwandes, der mittels der Bediengeschwindigkeit der Station in den erforderlichen Zeitaufwand umgerechnet wurde; vgl. Abschn. 2.2); damit war auch die Angabe von Bedienumfängen in Prozeßmustern festgelegt auf deren Zeit- bzw. Arbeits-Aufwand. Wir engten ferner die Reihenfolgevorschriften über den einzelnen Bedienwünschen einer Lasteinheit auf den streng sequentiellen Fall ein (was uns überhaupt erst berechtigte, über "Kunden" als Repräsentanten von Lastprozessen zu reden).

Beschreibung der Bedienwünsche

Der Besuch einer Station durch einen einzelnen Kunden gilt der Erledigung einer ganz bestimmten Arbeit von (im Prinzip) exakt bestimmbarem Umfang w (ein Prozessor soll eine Folge von w Anweisungen seiner Maschinensprache abarbeiten; über eine Leitung soll eine Nachricht der Länge w bits sequentiell übertragen werden). Andere Kunden, welche dieselbe Station benötigen, werden i.allg. Bedienwünsche anderen Umfangs aufweisen. Wir müßten also, wollten wir (im deterministischen Sinne) quantitativ präzise arbeiten, zur konkreten Beschreibung der Anforderungen an die betrachtete Station die Bedienwünsche aller Kunden (bei allen möglichen Betriebsabläufen!) erfassen. Dies übersteigt offen-

sichtlich unsere Fähigkeiten: Schon aufwandsmäßig (auch wenn wir zu Zwecken der prinzipiellen Durchführbarkeit auf die Unbeschränktheit der Menge betrachteter Betriebsabläufe verzichten würden), aber auch aus Gründen unseres beschränkten Wissens über die (bevorstehenden, u.U. Eingabedaten-abhängigen) Bedienwünsche der Kunden. Als Ausweg bietet sich an, anstatt der einzelnen Bedienwünsche aller Kunden, die Gesamtheit der Bedienwünsche aller Kunden gemeinsam zu charakterisieren: Wir akzeptieren, daß wir über Einzel-Bedienwünsche keine deterministisch präzisen Angaben zu machen vermögen, setzen aber voraus, daß wir über die Gesamtheit der Bedienwünsche insofern Kenntnisse besitzen, als wir wissen, mit welcher Häufigkeit sie "groß" oder "klein" sind, einen bestimmten Wert annehmen, in ein bestimmtes Werteintervall fallen (im konkreten Fall könnten wir diese Kenntnisse ihrerseits auf Messungen abstützen, aus einer Studie der Häufigkeiten von Eingabedaten-Kombinationen beziehen, aber auch - sinnvoll - hypothetisch annehmen). Das Ihnen bekannte mathematische Modell für eine solche Sachlage ist die Zufallsvariable: Der Umfang des einzelnen Bedienwunsches wird festgelegt durch eine (eindimensionale, numerische) Zufallsvariable W, ihrerseits charakterisiert etwa durch ihre Verteilungsfunktion $FW(w) := P[W \leq w]$. Mehr noch (und weiter einschränkend!): Die Folge der auf eine Station zukommenden Bedienwünsche wird festgelegt durch eine Folge von Zufallsvariablen $W_1, W_2, ...$, die als unabhängig identisch verteilt angesehen werden.

Zufallsvariable

Beschreibung der Bedienreihenfolge

Nach der Festlegung der Bedienwunsch-Umfänge (für jede Netzstation) gilt es nun, die seitens der Lastprozesse geforderte Bedien-Reihenfolge (die Folge der besuchten Stationen) zu charakterisieren. Wir verwenden eine Überlegung analog der bei den Bedienwünschen: Zwar mag es im konkreten Einzelfall des Besuches einer Station i durch einen Kunden zu aufwendig sein (oder unsere Kenntnisse übersteigen), die nach Abfertigung an Station i als nächste zu besuchende Station j deterministisch festzulegen. Insgesamt über alle Kundenbesuche an Station i aber setzen wir Kenntnisse voraus, die aussagen, mit welcher Häufigkeit nach einer Station i eine Station j, eine Station k, u.s.f., angefordert wird. Im mathematischen Modell setzen wir für diesen Zweck eine diskrete Zufallsvariable mit Wertebereich = Menge aller Netzstationen ein und legen (einschränkend!) fest, daß diese Zufallsvariable ihre Werte (den folgenden Besuch bezeichnend) mit festen (vom übrigen Geschehen unbeeinflußten) Wahrscheinlichkeiten einnimmt. Wir nennen diese Wahrscheinlichkeiten "Wechselwahrscheinlichkeiten" ("routing probabilities") und notieren sie unter Berücksichtigung aller Stationen mit Hilfe einer "Wechselmatrix" H, deren Elemente h_{ij} angeben, mit welcher Wahrscheinlichkeit auf einen Besuch an Station i einer an Station j folgt.

Zufallsvariable

Wechselwahrscheinlichkeiten, Wechselmatrix

Beschreibung eines Prozeßmusters

Mit den Bedienwünschen für alle Stationen und den Wechselwahrscheinlichkeiten zwischen allen Stationen ist ein Prozeßmuster vollständig charakterisiert. Verbleibt zu klären, wie einzelne Prozesse (die sich gemäß dieses Prozeßmusters verhalten) entstehen. Wie in Abschn. 2.1 schon ausgeführt, unterscheiden wir zwei alternative Fälle:

temporäre Prozesse

- aus der Umgebung generierte, d.h. temporäre, Prozesse - über deren Generierungsvorschriften wir uns noch Klarheit zu verschaffen haben. Wir nehmen dazu (einschränkend!) an, daß temporäre Prozesse je einzeln initiiert werden (also nicht mehrere zugleich) und folgen dem inzwischen bewährten Muster: Zwischen je zwei aufeinanderfolgenden Prozeßinitiierungen verstreicht eine gewisse Zeit, der sog. "Zwischenankunftsabstand". Diese Zeit wird nicht im Einzelfall deterministisch beschrieben, sondern insgesamt festgelegt durch die Modellvorstellung, daß die Folge der Ankunftsabstände durch eine Folge un-

Ankunftsabstände

abhängig identisch verteilter Zufallsvariabler $A_1, A_2, \ldots$ repräsentiert ist, etwa charakterisiert durch die Verteilungsfunktion FA(a). Das diese temporären Prozesse beschreibende Prozeßmuster muß vervollständigt werden einerseits durch Wechselwahrscheinlichkeiten, welche einem neu initiierten Prozeß die erste zu besuchende Station zuweisen und solche, die ein (endgültiges) Verlassen des Stationsnetzes von (u.U. verschiedenen) Stationen aus zulassen. Wie in Abschn. 2.2 fassen wir dazu die Umgebung als Pseudostation mit ausgezeichnetem Index 0 auf und reichern die Wechselmatrix H mit Größen h_{0i} (Wahrscheinlichkeit, daß auf Systemeintritt Station i folgt)und h_{j0} (Wahrscheinlichkeit, daß nach Station j Systemabgang folgt) an.

System -eintritt -austritt

- dauernd existierende, d.h. permanente, Prozesse - die nie generiert werden, aber in einer zu spezifizierenden festen Anzahl, sagen wir: N, vorhanden sind.

permanente Prozesse

Bevor wir die damit abgeschlossene Charakterisierung der Last-Komponente eines stochastischen Verkehrsnetzes nochmals zusammenfassen, sei auf eine Reihe von Vereinfachungen hingewiesen, die wir implizit getroffen haben, und die später wieder fallengelassen werden sollen. Zum einen haben wir uns auf ein einziges Prozeßmuster festgelegt, dem alle betrachteten Prozesse folgen. Dies ist der sog. "Ein-Ketten-" (single chain) Fall. Es mag aus Gründen der Realitätsnähe durchaus naheliegen, mehrere unterschiedliche Prozeßmuster vorzusehen (eines für Stapelaufträge, eines für Datenbank-Transaktionen, ...), in welchem Fall die obigen Beschreibungen zu vervielfachen wären; wir werden darauf zurückkommen. Zum anderen wurden die u.U. mehrfachen Besuche einer Station durch denselben Kunden nicht voneinander unterschieden (gleiche Bedienwunschverteilung, gleiche Wechselwahrscheinlichkeiten), obwohl auch hier ein Bedürfnis größerer Realitätsnähe vorliegen könnte (z.B. von 3 Besuchen ist der erste "lang", die beiden anderen "kurz"); diesem Bedürfnis kann man durch Einteilung der Besuche einer Station in verschiedene "Klassen" beikommen samt der Möglichkeit, daß Kunden ihre Besuchsklasse von Besuch zu Besuch ändern. Zunächst bleiben wir bei unserem "Ein-Klassen-" (single class) Fall und kommen später auf die angedeutete Verallgemeinerung zurück. Zusammenfassend:

implizite Vereinfachungen

eine Kette

eine Klasse

Annahmen 2.3.1: Bezeichne I die Menge der Stationen eines Verkehrsnetzes gemäß Abb. 2.1.1. Die Last-Komponente des Verkehrsnetzes wird im stochastischen Ein-Ketten-Fall beschrieben durch ein einzelnes Prozeßmuster sowie entweder (beim offenen Modell) durch eine Prozeßinitiierungsvorschrift oder (beim geschlossenen Modell) durch die Zahl N vorhandener Prozesse. Das Prozeßmuster wird im Ein-Klassen-Fall beschrieben durch eine Menge von (Bedienwunsch-) Variablen W_i; i I und eine (Stations-Wechsel-)Matrix der Dimension $|I|+1$ (beim offenen Modell) bzw. $|I|$ (beim geschlossenen Modell). Dabei ist verabredet, daß die Folge der Bedienwünsche für jede Station i durch eine Folge unabhängiger Zufallsvariabler repräsentiert ist, die identisch verteilt sind gemäß der Verteilung von W_i. Ferner, daß die Elemente h_{ij} der Matrix H bedingte Übergangswahrscheinlichkeiten sind, welche die Wechselwahrscheinlichkeiten Station i - Station j festlegen; beim offenen Modell dient ein besonderer Bezeichner (z.B. "0", 0 I) der Benennung der Umgebung und die Matrixelemente h_{0i} bzw. h_{j0} der Festlegung der Verzweigungswahrscheinlichkeiten zu Beginn eines temporären Prozesses bzw. seiner (bedingten) Beendigungswahrscheinlichkeiten. Die Prozeßinitiierungsvorschrift des offenen Modells ist festgelegt durch eine (Zwischenankunfts-)Variable A und die Verabredung, daß Prozesse einzeln und mit unabhängig und identisch (gemäß der Verteilung der Zufallsvariablen A) verteilten Ankunftsabständen initiiert werden.

Last-Komponente

Stations-verhalten

Wir haben uns bis jetzt ausschließlich um die Last-Komponente stochastischer Verkehrsnetze gekümmert. Was ist zu der Maschinen-Komponente zu sagen, d.h. zu der Menge I der Stationen? Erinnern wir uns an Abschn. 2.2 und daran, daß als wesentlichste Charakteristik einer Station ihre Bediendisziplin zu beachten ist; also zu beschreiben ist, unter welchen Umständen welcher anwesende Kunde mit welcher Geschwindigkeit bedient wird (bzw. allgemeiner, welche anwesenden Kunden mit welchen Geschwindigkeiten bedient werden). Über die konkreten Bediendisziplinen müssen wir für die jetzigen Überlegungen auch gar nicht viel mehr aussagen. Die allermeisten bedienen sich eines rein deterministischen Mechanismus und arbeiten mit rein (stations-) lokalen Informationen, d.h. auf der Basis eines lokalen "Zustands". Mit den Beispielen der Testfrage 2.2.9:

Bedien-disziplin

lokaler Zustand

z.B. FCFS

- Die FCFS-Disziplin "weiß" (d.h., hält als Zustand fest), wieviele Kunden anwesend und in welcher Reihenfolge sie eingetroffen sind. Sie bedient unter den Anwesenden jeweils den Erstankömmling vollständig (d.h. entsprechend des Gesamtumfangs seines Bedienwunsches) und zwar als Station von konstanter Bedienkapazität mit einer festen, zustandsunabhängigen Geschwindigkeit von r(AE/ZE).

z.B. HOL

- Die HOL-Disziplin kennt (als lokalen Zustand) die Anzahl der Anwesenden jeder Prioritätsklasse sowie die Reihenfolge ihrer Ankunft. Sie bedient aus der höchsten besetzten Prioritätsklasse deren Erstankömmling mit zustandsunabhängiger Geschwindigkeit.

Testfrage

Testfrage 2.3.2: Erklären Sie auf die soeben geübte Art Zustand und Bediengeschwindigkeiten der RR-Disziplin (vgl. Testfrage 2.2.9) und der Verzögerungsstation (vgl. Abb. 2.2.10 und zugehöriger Text).

räumliche Kapazität

Eine Zusatzangabe ist für die von uns betrachteten Stationen vonnöten: Wir setzen bis auf weiteres voraus, daß sie von unbeschränkter (räumlicher) Kapazität sind, d.h. daß sie (im Modell!) unbegrenzt viele Kunden aufnehmen können. Diese Voraussetzung ist wesentlich, da ohne sie auf das Ende der Bedienung eines Kunden an einer Station i und seines (auf der Basis seines Prozeßmusters) daran anschließend geäußerten Wechselwunsches in Richtung Station j nicht unbedingt ein Wechsel nach Station j erfolgen würde - Station j könnte ja, bei begrenzter Kapazität, zu diesem Zeitpunkt voll belegt sein. Solche "Blockierungen" der Kundenbewegung (die durchaus nicht unrealistisch sind!) wollen wir im Moment von unseren Betrachtungen ausschließen.

Blockierung

Testfrage

Testfrage 2.3.3: Warum mußten wir uns bei der betriebsanalytischen Interpretation der Verkehrsnetze (s. Abschn. 2.2) nicht um begrenzte Kapazitäten von Stationen kümmern, bzw. bei welchen der betriebsanalytischen Untersuchungen wäre es wesentlich, sich doch über Kapazitätsbegrenzungen Gedanken zu machen?

Fassen wir analog Ann. 2.3.1 bzgl. der Last-Komponente eines stochastischen Verkehrsnetzes nun auch unsere Voraussetzungen hinsichtlich der zugehörigen Maschinen-Komponente zusammen:

Maschinen-Komponente

Annahmen 2.3.4: Die Maschinen-Komponente eines stochastischen Verkehrsnetzes besteht aus einer endlichen Menge I von Bedienstationen. Wir bezeichnen gelegentlich die Mächtigkeit der Menge mit M(=|I|). Eine spezielle Vernetzungsstruktur der Stationen ist nicht vorgegeben (Sie können sich z.B. eine vollständige Vernetzung - von jeder Station zu jeder Station - vorstellen oder auch - aus

graphisch-darstellerischen Gründen - die Vernetzungsstruktur, die sich aus den Elementen $h_{ij} \neq 0$ der Wechselmatrix der Last-Komponente ergibt). Alle Stationen sind von unbegrenzter räumlicher Kapazität. Jeder Station zugeordnet ist ihre spezifische Bediendisziplin. Jede Bediendisziplin arbeitet auf einem geeignet angelegten Zustandsraum. Die Station nimmt Zustände (aus diesem Zustandsraum) ausschließlich auf der Basis lokaler (in der Station beobachtbarer) Umstände ein, so daß ein Stationszustand auch als kompakte Darstellung der gesamten Vergangenheit der Station gesehen werden kann; kompakt in dem Sinne, daß alle für das Arbeiten der Disziplin unwesentlichen Umstände "vergessen" werden können. Die Wirkung einer Disziplin besteht darin, daß sie jedem (aufgrund von Geschehnissen der Vergangenheit) anwesenden Kunden eine (momentane, für die Dauer dieses Zustands bestehende) Bediengeschwindigkeit zuweist; der Fall der (momentanen) Nicht-Bearbeitung eines Kunden ist durch "Bediengeschwindigkeit=0" mit erfaßt.

stochastische Verkehrsnetze

Wir waren zu Beginn des Abschnitts mit dem Ziel gestartet, die Regeln und Gesetze festzulegen, nach denen sich die Betriebsabläufe in einem Verkehrsnetz (als Leistungsmodell eines Rechensystems) vollziehen. Wir haben dieses Ziel (in den Annahmen 2.3.1 und 2.3.4) durch eine für die jetzigen Zwecke hinreichend präzise Spezifikation stochastischer Verkehrsnetze erreicht. Der nächste Schritt muß offensichtlich darauf gerichtet sein, ein solches stochastisches Modell zu analysieren, d.h. aus der quantitativen Konkretisierung eines stochastischen Verkehrsnetzes mittels geeigneter analytischer Techniken auf die interessierenden Leistungsgrößen (Durchsätze, Verweilzeiten, etc.) zu schließen.

Analyse

Resultate

Welche Art von Resultaten können wir überhaupt erwarten? Konzentrieren wir uns (als Beispiel) auf den Durchsatz einer bestimmten Station. Sei dieser (in Anlehnung an Abschn. 2.2) definiert als Zahl der Kundenabgänge von dieser Station pro Zeiteinheit. Betrachten wir diese Größe für einen einzelnen Betriebsablauf (wie er von unserem stochastischen Verkehrsnetz erzeugt werden könnte), indem wir über die Zeitachse eine Folge aneinander anschließender Zeitintervalle jeweils der Länge 1ZE legen:

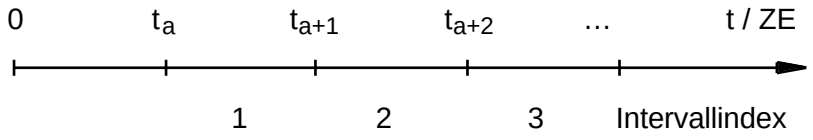


Abbildung 2.3.5a: Strukturierung der Zeitachse

Durchsätze

Notieren wir jetzt die Durchsätze in jedem dieser Zeitintervalle, dann erhalten wir (im Prinzip) ein Ergebnis entsprechend Abb. 2.3.5b. Betrachten wir einen zweiten Betriebsablauf (auch dieser sei von unserem stochastischen Verkehrsnetz erzeugt worden, das ja alle möglichen Betriebsabläufe beschrieb), dann erhalten wir bei gleicher Strukturierung der Zeitachse zwar das gleiche prinzipielle Bild, aber möglicherweise mit anderen Durchsatzwerten, z.B. mit denen der Abb. 2.3.5c. Diese Unterschiedlichkeit der Beobachtungen entspricht sicher den Erwartungen, waren doch Bedien- und Stationswechsel-Wünsche der Kunden nur stochastisch festgelegt, konnten also von Betriebsablauf zu Betriebsablauf variieren - und damit die resultierenden Durchsätze ebenfalls. Wir können also nicht erwarten, den Durchsatz einer Station in einem bestimmten Zeitintervall

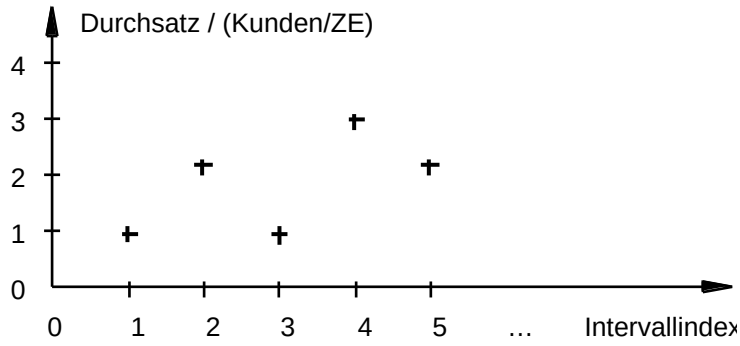


Abbildung 2.3.5b: Durchsatz einer Station bei einem Betriebsablauf

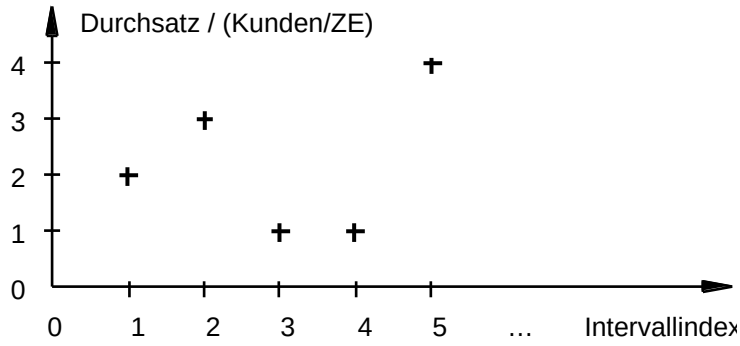


Abbildung 2.3.5c: Durchsatz einer Station bei einem zweiten Betriebsablauf

als deterministische Größe errechnen zu können. Vielmehr müssen wir erneut auf eine probabilistische Charakterisierung aus sein: Der Durchsatz im i-ten Intervall wird durch eine Zufallsvariable, nennen wir sie D_i, beschrieben; wir wissen über die ZV D_i Bescheid, wenn wir ihre Verteilung angeben (ausrechnen) können. Damit nicht genug: Wir müssen darauf gefaßt sein, daß bei der Beobachtung eines einzelnen Betriebsablaufes verschiedene Durchsatzwerte, z.B. d_i und d_{i+1}, voneinander abhängen - etwa in dem Fall, daß die Bedienwünsche der Kunden in 80% aller Fälle einer Bearbeitungszeit von mehr als 2ZE entsprechen, so daß (unter der Annahme nicht-unterbrechender Bedienung) aus $d_i>0$ in mehr als 80% der Fälle $d_{i+1}=0$ folgt. Entsprechend müssen wir darauf eingestellt sein, daß die ZV D_i und D_{i+1} nicht unabhängig sind, ja daß auch Abhängigkeiten zwischen weiteren Durchsatz-ZV bestehen und daher eine vollständige Charakterisierung des Durchsatzes einer Station der Angabe der gemeinsamen Verteilung der Menge von Zufallsvariablen D_i; i=1,2,... bedarf. Wir können dem Kind jetzt auch den richtigen Namen geben: Die Familie von Zufallsvariablen (D_i; i=1,2,...) ist ein Beispiel eines "stochastischen Prozesses", im konkreten Falle eines mit diskretem "Zustandsraum" (jedes D_i kann nur Werte aus der abzählbaren Menge $\mathbf{N_0}$ annehmen) und diskreter "Parametermenge" (i durchläuft die abzählbare Menge **N**).

stochastischer Prozeß

Wir wollten herausarbeiten, welche Art von Resultaten wir aus der Analyse stochastischer Verkehrsnetze erwarten können. Für die Resultatgröße "Durchsatz einer Station" haben wir erkannt, daß das vollständige Resultat durch die Charakterisierung des stochastischen Prozesses (D_i; i=1,2,...) gegeben wäre. Aus dieser Charakterisierung könnten wir ersehen, welche Folgen von Durchsatzwerten (d_1,d_2,d_3,...) - jede derartige Folge ist genau einem Betriebsablauf zugeordnet -

überhaupt auftreten und mit welcher Wahrscheinlichkeit (für die Realität interpretiert: mit welcher relativen Häufigkeit) solche Folgen bestimmte Eigenschaften aufweisen. Um Sie an der potentiellen Fülle der hier angesprochenen Information nicht von vornherein verzweifeln zu lassen: Wir können bei der geforderten Charakterisierung des stochastischen Prozesses natürlich bescheiden sein; vielleicht interessieren uns gar nicht die individuellen Folgen ($d_1,d_2,d_3,...$) sondern nur die Mittelwerte der Beobachtungen ($\bar{d}_1,\bar{d}_2,\bar{d}_3,...$), quer über alle Betriebsabläufe gesehen, so daß statt der vollständigen Charakterisierung des stochastischen Prozesses (D_i; i=1,2,...) nur die Folge der Erwartungswerte ($E[D_1],E[D_2],...$) "auszurechnen" wäre. Vielleicht auch ist der angesprochene Prozeß unter bestimmten Bedingungen so "gutartig", daß die Verteilungen der einzelnen Durchsatzvariablen D_i sich nicht unterscheiden, so daß auch ihre Erwartungswerte zusammenfallen und wir von (einem einzigen) "mittleren Durchsatz" reden können (womit wir auf einer Betrachtungsebene landen, die den betriebsanalytischen Untersuchungen des Abschn. 2.2 stark ähnelt). In der Tat werden solche "stationär" genannten Prozesse im Zentrum unseres Interesses stehen - womit unsere Aufgabe vielleicht nicht mehr ganz so umfangreich aussieht.

Analyseaufwand

stationäre Prozesse

Stochastische Prozesse werden uns bei unserer Analyse auf Schritt und Tritt begegnen. Nehmen wir uns statt des zuvor betrachteten Beispiels "Durchsatz einer Station" ein anderes vor: "(Kunden-) Verweilzeit in einer Station", definiert als Länge des Zeitintervalls zwischen der Ankunft eines Kunden an der Station und seinem Abgang von der Station. Betrachten wir diese Größe wieder für einen einzelnen Betriebsablauf und beginnen unsere Beobachtung zu irgendeinem Zeitpunkt t_a. Der erste Kunde, der nach diesem Zeitpunkt die Station verläßt, habe die Zeit t_1 in der Station verbracht, hat also eine Verweilzeit t_1, der nächste eine Verweilzeit t_2, usw. Insgesamt erhalten wir eine Folge von Verweilzeiten (t_i; i=1,2,...). Betrachten wir einen zweiten Betriebsablauf, beginnen wieder zum Zeitpunkt t_a mit der Beobachtung, so erhalten wir wieder eine Folge von Verweilzeiten (t_i'; i=1,2,...), aber aller Voraussicht nach (wegen der Stochastik im Erzeugungsmechanismus "stochastisches Verkehrsnetz") mit von der ersten Folge abweichenden Werten. Wir wissen bereits Abhilfe für die dadurch entstehende Schwierigkeit bei der Charakterisierung der Folge: Jede der Beobachtungen wird als Zufallsvariable T_i (i=1,2,...) aufgefaßt, die Menge der Beobachtungsfolgen als stochastischer Prozeß (T_i; i=1,2,...); wieder müssen wir von einer gegenseitigen Abhängigkeit der T_i ausgehen (so ist es hochwahrscheinlich, daß bei einer FCFS-Station auf ein "großes" t_i ein "ziemlich großes" t_{i+1} folgt: Nachdenken!); wieder müßte eine (vollständige) Charakterisierung der Zufallsvariablen-Folge (T_i; i=1,2,...) auf die Angabe der gemeinsamen Verteilung aller beteiligten ZV hinauslaufen.

Verweilzeit

Allerdings sollte eines klar sein: So interessant eine geeignete Charakterisierung der stochastischen Prozesse (D_i) bzw. (T_i) auch als Resultat der Analyse erscheinen mag, keiner der Prozesse bietet sich als Grundlage der Analyse an; die Information, die in Feststellungen der Art "D_i hat den Wert 3" oder "D_i liegt in 80% der Fälle über 2" enthalten ist, ist einfach zu gering, als daß wir (deterministisch oder probabilistisch) auf die Werte anderer D.'s schließen könnten; bzw. globaler betrachtet: Unser sorgsam entworfener Generierungsmechanismus "stochastisches Verkehrsnetz" ist in den Prozessen (D_i) und (T_i) gar nicht reflektiert und eine isolierte Betrachtung dieser Prozesse daher aussichtslos.

Sehen wir uns unseren Generierungsapparat nach Ann. 2.3.1/2.3.4 nochmals an: Jede Station nahm im Verlaufe eines Betriebsablaufs unterschiedliche Zustände aus ihrem spezifischen (lokalen) Zustandsraum ein. Der Stationszustand diente

Analysegrundlage

in den vorgenannten Überlegungen dazu, den anwesenden Kunden (nach Maßgabe der Bediendisziplin) eine Bediengeschwindigkeit zuzuordnen, ihnen also einen Bedienungsfortschritt (eine Reduktion ihres Bedienwunsches bzw. des nach bereits erfolgter teilweiser Bedienung verbliebenen Bedienwunsch-Restes) zukommen zu lassen. Betrachten wir eine Station i während eines Betriebsablaufs, dann stellen wir zu einem Zeitpunkt t fest, daß sie sich in einem bestimmten Zustand befindet, nennen wir ihn $z_i(t)$. (Welche Information in z_i enthalten ist, müssen wir im Moment nicht konkret festlegen - dazu kommen wir aber in Kap. 3). Bei Beobachtung eines zweiten Betriebsablaufs wird sich (gleiche Station, gleicher Zeitpunkt) ein Zustand $z_i'(t)$ finden, wobei i.allg. $z_i(t) \neq z_i'(t)$; das vertraute Bild demnach: Der Zustand der Station i zum Zeitpunkt t muß durch eine Zufallsvariable $Z_i(t)$ charakterisiert werden. Weiter: Entlang eines einzelnen Betriebsablaufs werden Stationszustände zu verschiedenen Zeitpunkten, z.B. $z_i(t)$ und $z_i(t')$, voneinander abhängen (für "kleine" Zeitdifferenzen t'-t werden wahrscheinlich auch die zugehörigen Zustände "nur wenig" voneinander abweichen); demzufolge sind auch die Zufallsvariablen $Z_i(t)$ und $Z_i(t')$ als abhängig anzusehen. Bezeichnen wir die Menge aller betrachteten Zeitpunkte (aller Zeitpunkte also, zu denen uns der Stationszustand interessiert) mit T, dann haben wir in $(Z_i(t); t \in T)$ erneut einen stochastischen Prozeß vor uns, dessen vollständige Charakterisierung durch Angabe der gemeinsamen Verteilung aller $Z_i(t)$, $t \in T$, erfolgen müßte. Erschwerend kommt hier allerdings dazu, daß wir die "Zeit" als kontinuierliche Größe auffassen werden, also T eine (überabzählbare, konvexe) Untermenge der reellen Zahlen darstellt: $(Z_i(t); t \in T)$ ist ein stochastischer Prozeß mit "kontinuierlicher Parametermenge". Und: Welcher Art der Zustandsraum ist (diskret, kontinuierlich, gemischt), darüber müssen wir erst noch diskutieren.

Stationszustand

Zufallsvariable

stochastischer Prozeß

Ein Schritt noch: Die Zustände der verschiedenen Stationen im Netz, also die $z_i(t)$, $i \in I$, sind natürlich ebenfalls voneinander abhängig (denken Sie z.B. an ein geschlossenes Modell, bei dem die Aufteilung der Kunden auf die verschiedenen Stationen zwar mit der Zeit variiert, aber eben unter der Randbedingung, daß die Gesamtzahl konstant bleibt). Dies macht es notwendig, statt der Betrachtung der Zustände einzelner Stationen den Gesamtzustand des Netzes zu beobachten, etwa in Form des "Zustandsvektors" $\underline{z}(t) = (z_i(t); i \in I)$. Diesen gilt es wegen der stochastischen Schwankungen wieder als (hier: vektorielle) Zufallsvariable $\underline{Z}(t) = (Z_i(t); i \in I)$ aufzufassen und über der Zeit als stochastischen Prozeß $(\underline{Z}(t); t \in T)$ zu studieren.

Systemzustand

Wir haben damit zwar ein Studienobjekt beträchtlicher Komplexität konstruiert, dafür aber auch eines, das wir (im Gegensatz zu den simplen Prozessen (D_i) bzw. (T_i)) mit zumindest prinzipieller Hoffnung auf Erfolg untersuchen können. Diese Hoffnung gründet sich auf die Tatsache, daß die Veränderung des Modellzustands über der Zeit direkt unserem (bekannten!) Generierungsmechanismus unterliegt. Denken wir darüber kurz noch einmal nach, und zwar der einfacheren Vorstellungsmöglichkeit halber in deterministischer Terminologie (eine Übertragung in die Welt der Stochastik müssen wir anschließend sicherlich vornehmen): Solange kein Kundenzugang erfolgt, arbeitet jede Station autonom vor sich hin; die anwesenden Kunden werden gemäß Bediendisziplin bearbeitet, evtl. zu bestimmten Zeiten unterbrochen, verlangsamt, beschleunigt usw.; der lokale Zustand jeder Station verändert sich also in festgelegter (und vorhersehbarer!) Weise; jeder Kunde hatte die Station mit einem bekannten Bedienwunsch betreten, wir hatten ihm bekannte (evtl. wechselnde) Bearbeitungsgeschwindigkeiten zukommen lassen, es läßt sich absehen (ausrechnen!), wann er fertig bedient ist; zu diesem Zeitpunkt schließt sich i.a. an den soeben absolvierten Bedienwunsch ein nächster an einer anderen (bekannten) Station an; der Kunde

Objekt der Analyse

wechselt zu dieser Station, der Globalzustand des Netzes verändert sich entsprechend (bzw.: der Lokalzustand von Quell- und Zielstation ändert sich simultan); die Stationen arbeiten wieder eine Weile autonom; dann erfolgt ein Kundenwechsel, usw. - die Zustandsübergänge sind also sämtlich nachvollziehbar (und prognostizierbar), wir haben die Dynamik des Modells (im Prinzip) voll im Griff. In stochastischer Terminologie ist ein Bedienwunsch natürlich nicht zu einem festen Zeitpunkt abgearbeitet, sondern mit einer gewissen Wahrscheinlichkeit in einem bestimmten Zeitintervall; wechselt ein Kunde nicht zu einer festen Folgestation, sondern mit gewissen Wahrscheinlichkeiten zu bestimmten Folgestationen; unsere Aussagen werden unschärfer, an der prinzipiellen Nachvollziehbarkeit bzw. Prognostizierbarkeit ändert das nichts. Es wird Gegenstand des nächsten Kapitels sein, das somit aufgestellte Analyseprogramm konkret auszufüllen und zu sehen, in welchen Fällen wir uns von der prinzipiellen Analysierbarkeit des stochastischen Zustandsprozesses ($\underline{Z}(t); t \in T$) zu effektiven Resultaten durcharbeiten können.

LEERSEITE

Literatur zu Kap. 2 und allgemein zu "operational analysis"

BrDe82 Brumfield,J.A./Denning,P.J.; Error analysis of homogeneous mean queue and response time estimators; Performance Evaluation Review vol.11(1982/83) nr.4

Brya80 Bryant,R.M.; On homogeneity in M/G/1 queuing systems; in: Proc. PERFORMANCE ´80: Performance Evaluation Rev. vol.9(1980) nr.2 pp.199-208

Buze71 Buzen,J.P.; Analysis of system bottlenecks using a queuing network model; in: Proc. ACM/SIGOPS workshop on system performance evaluation; ACM 1971

Buze76 Buzen,J.P.; Fundamental operational laws of computer system performance; Acta Informatica vol.7(1976) nr.2

Buze78 Buzen,J.P.; Operational analysis: An alternative to stochastic modelling; in: Ferrari,D.; Performance of Computer Installations; North Holland 1978

ChLa74 Chang,A./Lavenberg,S.S.; Work rates in closed queueing networks with general independent servers; Operations Research vol.22(1974) nr.4

DeBu77 Denning,P.J./Buzen,J.P.; Operational analysis of queueing networks; in: Proc. PERFORMANCE ´77: Beilner/Gelenbe (eds); Measuring, modelling and evaluating computer systems pp.151-172; North Holland 1977

DeBu78 Denning,P.J./Buzen,J.P.; The operational analysis of queueing network models; Computing Surveys vol.10(1978) nr.3 pp.225-261

DeKo82 Denning,P.J./Kowalk,W.; Error analysis of the mean busy period of a queue; in: Proc. 10th IMACS World Congress on System Simulation and Scientific Computation, 1980

Denn81 Denning,P.J. ; Performance analysis: Experimental computer science at its best; CACM vol.24(1981) nr.11 pp.725-727

Dörf78 Dörfler, W.; Mathematik für Informatiker, Band 2: Methoden aus der Analysis; C. Hanser, 1978

Gele82 Gelenbe,E.; Stationary deterministic flows in discrete systems I; Performance Evaluation Review vol.11(1982) nr.4 p.89ff

Kowa80 Kowalk,W.; Error analysis in operational queueing models; in: 10th IMACS World Congress on System Simulation and Scientific Computation vol.4, 1980

Kowa81 Kowalk,W.; Conservation laws in operational analysis; in: Proc. PERFORMANCE ´81: Kylstra (ed); North Holland 1982

Kowa82 Kowalk,W.; Verkehrsanalyse in endlichen Zeiträumen; Springer, IFB vol.55(1982)

Kowa83 Kowalk,W.; An operational view on renewal theory; in: Performance Evaluation Review special issue 1983

Kowa83 Kowalk,W.; Erweiterte Methoden der operationalen Analyse; Angewandte Informatik vol.25(1983) nr.1

Kowa89 Kowalk W.; Operationale Wartetheorie; Springer IFB vol.196 1989

LZGS84 Lazowska,E.D./Zahorjan,J./Graham,G.S./Sevcik,K.C.;
Quantitative system performance - Computer system analysis using queueing network models; Prentice Hall 1984

MoBr88 Molnar R., Bruell S.C.; Sensitivity analysis of operational formulae;
J. Computer Systems, Science and Engineering vol.3(1988) nr.2 pp.51-66

Rood79 Roode,J.D.; Multiclass operational analysis of queueing networks;
in: Proc. PERFORMANCE ´79: Arato et al (eds); Performance of computer systems;
North Holland 1979

ThBA85 Thareja,A.K./Buzen,J.P./Agrawal,S.C.; BEST/1-SNA: A software tool for modelling and analysis of IBM SNA networks; in Potier,D. (ed): Modelling techniques and tools for performance analysis, North Holland 1985 pp.81-98

Wu82 Wu,L.T.; Operational models for the evaluation of degradable computing systems;
Performance Evaluation Review vol.11(1982) nr.4 pp.179ff

Lösungshinweise zu den Testfragen

2.1.2:

Temporäre Prozesse modellieren das Verhalten von Aufgaben, die von der Umgebung an die Maschine gestellt werden. Im Normalfalle werden dabei, mit gewissen zeitlichen Abständen, immer wieder Prozesse eines Prozeßmusters generiert (z.B. die Aufgaben der Stapelverarbeitung repräsentierend); solche Aufgaben sind aber irgendwann "abgearbeitet". Wären sie dies nicht, dann würde sich das System mit in Bearbeitung befindlichen Aufgaben immer weiter "anfüllen". Für den Fall einiger weniger tatsächlich "immer" vorhandener Prozesse (z.B. Verwaltungsprozesse) setzt man die permanenten Prozesse ein - und unterschlägt ggf. die explizite Generierung (die ja nur den "Systemstart" nachbilden würde).

2.2.4:

Die Feststellung "Ereignis x tritt mit Rate y auf" impliziert die Vorstellung von y Ereignissen x je Zeiteinheit. Selbst wenn Sie klarstellen, daß aufgrund der Berechnungsvorschriften nur "im Mittel" y Ereignisse x pro Zeiteinheit generiert sein können, führt

- ein zu kleines T
  (Beispiel: In 1 Stunde geschehen insgesamt 10 Ereignisse, T hat die Länge 1 sec, in diese Sekunde fällt ein/kein Ereignis)

- starke statistische Unregelmäßigkeit
  (Beispiel: In 1 Stunde geschehen insgesamt 10 Ereignisse, T hat die Länge 1 h, alle 10 Ereignisse fallen in die ersten 10 sec von T)

zu (intuitiv) falschen Vorstellungen.

2.2.9:

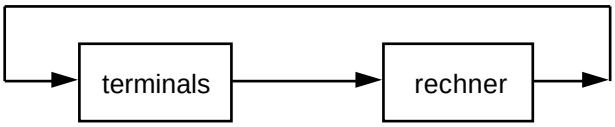
Sollten Ihnen die Bediendisziplinen nicht vertraut sein (in welchem Fall Sie unbedingt Ihre Betriebssystem-Kenntnisse auffrischen sollten!), hier eine kurze Charakterisierung:

- FCFS: Kunden werden in der Reihenfolge ihrer Ankunft, jeweils vollständig und ohne Unterbrechung, bedient;
- HOL: Jeder Kunde gehört einer aus einer festen Anzahl von (Prioritäts-) Klassen an; der Warteraum der Station enthält für jede Klasse eine gesonderte Warteschlange; die Bedienung jedes Kunden geschieht vollständig und ohne Unterbrechung; höhere Prioritätsklassen werden vor niedrigeren bedient, innerhalb jeder Prioritätsklasse gilt FCFS;
- RR: Der Warteraum besteht aus einer Warteschlange; Neuankömmlinge stellen sich am Ende der Schlange an; die Bedienung erfolgt entsprechend der Ordnung der Warteschlange, ist allerdings unterbrechend in dem Sinne, daß der Bediener einem Kunden maximal für die Dauer einer "Zeitscheibe" fester Länge verfügbar ist; ist die Bedienanforderung innerhalb dieser Zeitscheibe nicht zur Gänze abgewickelt, stellt sich der Kunde (mit seiner "Rest"-Bedienanforderung) wieder am Ende der Warteschlange an und wartet auf Fortsetzung der Bedienung u.s.f.

Zur Aufgabe: Machen Sie sich (durch Zeichnungen, Rechnungen o.ä.) klar, daß weder eine (beliebige!) Umordnung im Warteraum noch eine (beliebige!) Stückelung der Bedienung auf Gleichungen (2.2.7) einen Einfluß hat, solange nur "arbeitserhaltend" und "konstante Bedienkapazität" im Sinne von Def. 2.2.8 zutrifft, was hier insbesondere heißt, daß durch die Organisation der Disziplin keine Bedienkapazität verlorengeht: Also kein Aufwand für Sortieren, Unterbrechen, Wiederaufsetzen; kein Verlust bereits geleisteter Arbeit; keine "idle"-Zeiten trotz nichtleeren Warteraums.

2.2.11:

Alle Endgeräte werden zusammen durch eine Verzögerungsstation terminals repräsentiert, so daß sich folgende Maschine ergibt:



Wir benötigen nur noch ein Prozeßmuster:

```
LOOP
   terminals.denke;
   rechner.bearbeite
ENDLOOP
```

zu dem es n individuelle Prozesse gibt, jeder einem Benutzer (samt seinem nur noch gedachten individuellen Endgerät) zugeordnet. Die Umsetzung von "denke" in das physikalisch notierte "widme mir (z.B.) 20 sec" - die Denkzeit des Benutzers repräsentierend - bleibt unverändert (die u.U. zeitlich überlappend ablaufenden Denkvorgänge beeinflussen einander aufgrund der Verzögerungsstation-Fiktion in keiner Weise), eine Umsetzung von "bearbeite" ist, wie gehabt, nur bei weiterer Auflösung von rechner sinnvoll möglich.

2.2.28:

Die Aufgabe ist in ihrer Einfachheit fast trivial und auch ohne viel Formalismus mit gesundem Menschenverstand lösbar: Von den 31 cpu-Besuchen je Stapelauftrag endet der letzte mit einem Verlassen des Systems (nur von cpu aus ist Verlassen möglich), an die ersten 30 schließt sich je eine Anforderung an eine der E/A-Einheiten an; davon gelten 20 der disk_1, 10 der disk_2. Numerieren wir der Einfachheit halber die Stationen: disk_1 entspricht "1", disk_2 entspricht "2", cpu entspricht "3", dann erhalten wir bei einer Ankunftsrate der Stapelaufträge von c_0 (Systemankünfte/Zeiteinheit) die Durchsätze $c_3=31\cdot c_0$, $c_2=10\cdot c_0$, $c_1=20\cdot c_0$.

Bei einer komplexeren Aufgabe dieser Art ist es dagegen ratsam, formaler vorzugehen wie folgt:

$h_{03} = 1$	Jede Systemankunft geht nach cpu
$h_{13} = h_{23} = 1$	Jeder E/A-Abgang geht nach cpu
$h_{31} = 20/31$	
$h_{32} = 10/31$	Entsprechend obiger Überlegungen
$h_{30} = 1/31$	
$h_{ij} = 0$	für alle anderen (i,j)

Gleichungssystem (2.2.27b) erhält daher das Aussehen

$$
\begin{aligned}
c_0 &= h_{30}c_3 + 0 \\
c_1 &= h_{31}c_3 + 0 \\
c_2 &= h_{32}c_3 + 0 \\
c_3 &= h_{03}c_0 + h_{13}c_1 + h_{23}c_2 + 0
\end{aligned}
$$

Wegen der Homogenität des Gleichungssystems entfernen wir eine Gleichung (z.B. die letzte) und benutzen c_0 als Parameter

$$
\begin{aligned}
h_{30}c_3 &= c_0 \\
c_1 - h_{31}c_3 &= 0 \\
c_2 - h_{32}c_3 &= 0
\end{aligned}
$$

woraus sich die bereits bekannte Lösung

$$c_3 = \frac{1}{h_{30}} c_0 = 31\, c_0$$
$$c_1 = \frac{h_{31}}{h_{30}} c_0 = 20\, c_0$$
$$c_2 = \frac{h_{32}}{h_{30}} c_0 = 10\, c_0$$

ergibt.

2.2.32:

Gegenüber 2.2.28 ändert sich nur folgendes: Eine Umgebung (also "Station 0") gibt es nicht mehr, h_{03} und h_{30} fallen also weg. Dagegen gibt es einen neuen Weg cpu-cpu und die zugehörige Übergangshäufigkeit $h_{33}=1/31$ - dem Werte nach natürlich gleich dem alten h_{30}, da ja mit diesem Weg (zuvor und jetzt) die Beendigung eines Stapelauftrags angezeigt ist. Ziehen wir die Aufgabe formal durch (sie wäre nach wie vor einfach genug, um auch ohne Formalismus durchzukommen), dann haben wir es mit dem Gleichungssystem (2.2.31) zu tun:

$$h_{31}X_3 - X_1 = 0$$
$$h_{32}X_3 - X_2 = 0$$
$$h_{13}X_1 + h_{23}X_2 + (h_{33}-1)X_3 = 0$$

mit der zusätzlichen Gleichung ($j^*=1$ wegen Messung von c_1 an disk_1)

$$X_1 = 1$$

Lassen wir eine der oberen Gleichungen weg, z.B. die dritte, dann ergeben sich

$$X_3 = \frac{1}{h_{31}} X_1 = \frac{1}{h_{31}} = \frac{31}{20}$$
$$X_2 = h_{32} X_3 = \frac{h_{32}}{h_{31}} = \frac{1}{2}$$

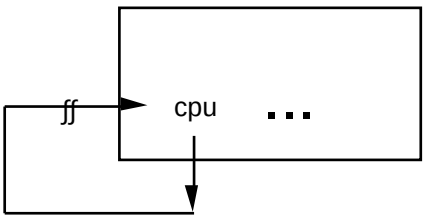
sowie nach (2.2.30) die Durchsätze (in Kunden/ZE)

$$c_2 = X_2 c_1 = \frac{40}{2} = 20$$
$$c_3 = X_3 c_1 = \frac{31\,40}{20} = 62$$

Die Abfertigungsrate für Stapelaufträge entspricht dem Verkehrsfluß cpu-cpu, hat also den Wert

$$c_3 \cdot h_{33} = 62/31 = 2$$

Stellen wir uns nun "in" den Rückkopplungsweg cpu-cpu und fassen das gesamte System als "Station" auf:



Wir kennen den Durchsatz (die Eingangsrate): $c_3 h_{33} = 2$ Kunden/ZE; die Zahl der Kunden im System: konstant gleich MP=5, d.h. auch im Mittel 5. Little's Gesetz (2.2.22b) ist anwendbar und er-

gibt als mittlere Bearbeitungszeit eines Stapelauftrags

$$\overline{V} = \frac{n}{a} = \frac{MP}{c_3 h_{33}} = \frac{5}{2} = 2.5\,ZE$$

2.2.34:

Der maximale Wert für c_0, nennen wir ihn $c_{0,max}$, ergab sich aus der Tatsache, daß für diesen Wert eine der Stationen, sagen wir Station k, einen Durchsatz $c_{k,max}$ erreichte derart daß, s. (2.2.30,2.2.33)

$$1 = b_{k,max} = \overline{B}_k \cdot c_{k,max}$$

diese Station also dauernd beschäftigt war. Würden wir nun c_0 über $c_{0,max}$, und damit c_k über $c_{k,max}$, hinaus steigern, dann könnte Station k die daraus resultierende Arbeit nicht mehr bewältigen: Der Warteraum von Station k würde sich fortlaufend weiter anfüllen mit Kunden, die auf Bearbeitung warten (real natürlich nur bis zu einer physikalisch gegebenen Grenze, die aber in unserem Modell nicht enthalten ist). Wie groß auch immer die Besetzung dieses Warteraums wird, an der Abgangsrate $c_{k,max}$ der Station k ändert sich nichts mehr: Die Station arbeitet ja bereits "mit voller Kraft". Da nun alle Durchsätze gemäß Verkehrsflußgleichgewicht (2.2.27) fest aneinander gebunden sind und Stationen $i \neq k$ auch die aus $c_k = c_{k,max}$ resultierenden Ankunftsraten c_i durchaus bewältigen, bleibt auch der Systemdurchsatz (jetzt am Ausgang des Systems als Abgangsrate System gemessen) an den Maximalwert $c_{k,max}$ gebunden. Insgesamt: Steigt die Ankunftsrate am System (Abgangsrate Umgebung) c_0 über den Wert $c_{0,max}$, dann nimmt die Abgangsrate vom System (Ankunftsrate Umgebung) a_0 den festen Wert $c_{0,max}$ an, das Verkehrsflußgleichgewicht ist gestört ($a_0 \neq c_0$); intern wächst die Kundenpopulation an Station k über alle Grenzen.

2.2.37:

Nein! Beschleunigen Sie nämlich eine Station i, die nicht die Flaschenhals-Station ist, dann sinkt dadurch zwar deren Auslastung

$$b_{i,neu} (= \overline{S}_{i,neu} \cdot X_i c_0) < b_{i,alt} (= \overline{S}_{i,alt} \cdot X_i c_o)$$

was aber auf den Durchsatz der Flaschenhals-Station keinerlei Einfluß hat (alle X. bleiben gleich!), somit auch nicht auf die Tatsache, daß die Flaschenhals-Station für $c_{0,max}$ voll ausgelastet ist. Wenn Sie eine Verbesserung des maximalen Systemdurchsatzes erreichen wollen, dann müssen Sie schon die Flaschenhals-Station beschleunigen: Trage diese Station den Index k, dann ist ja durch diese Maßnahme

$$\overline{S}_{k,neu} \cdot X_k < \overline{S}_{k,alt} \cdot X_k$$

erreicht, und der Wert $c_{0,max}$ ist auf einen höheren Wert festgelegt (sowohl wenn Station k Flaschenhals bleibt, als auch, wenn eine andere Station Flaschenhals wird; vgl. (2.2.33c) mit $j^*=0$).

2.3.2:

- Die RR-Disziplin "merkt sich" (als Zustand) alle anwesenden Kunden sowie eine lineare Ordnung über diesen Kunden. Wenn Kunden vorhanden sind, wird deren Erster (gemäß festgelegter Ordnung) mit fester Bediengeschwindigkeit r bedient - alle anderen warten, haben also Bediengeschwindigkeit 0. Wird der bediente Kunde während der zugestandenen Bedienzeitscheibe fertig, verläßt er die Station (bzgl. des lokalen Stationszustands wird er also "vergessen") - der (gemäß festgelegter Ordnung) Nächste ist jetzt Erster und wird mit Geschwindigkeit r bedient. Wird der Bediente innerhalb der Zeitscheibe nicht fertig, rutscht er "ans Ende" der linear ange-

ordneten Kunden (hat also Geschwindigkeit 0) - der Nächste rückt vor (bekommt also Geschwindigkeit r). Ein Neuankömmling an der Station stellt sich "am Ende" an.

- Die Verzögerungsstation kennt alle Anwesenden und weist jedem von ihnen eine konstante Bediengeschwindigkeit von z.B. r (AE/ZE) zu.

2.3.3:

Die Grundidee war ja die der Messung am System. Sollte also einer der Kunden (obwohl an Station i fertig bedient und mit dem Wunsch, nach Station j zu wechseln) aufgrund einer räumlichen Überlastung der Zielstation (Station j) nicht wechseln können, dann verbleibt er wohl an der Quellstation (Station i) und wird auch nicht als Abgang bzw. Ankunft gezählt: Die Messungen beinhalten bereits alle Blockierungseffekte. Andererseits: Wenn wir, von einer Messung ausgehend, zur Ermittlung von Flaschenhälsen hypothetisch Zugangsrate (beim offenen System) oder Kundenzahl (beim geschlossenen System) erhöhen, dann verfälschen Blockierungen unsere Resultate: So könnte z.B. eine Quellstation, die einen fertigen Kunden "nicht los wird", in ihrer weiteren Arbeit anhalten müssen: Die Station arbeitet nicht, obwohl u.U. Arbeit ansteht - eine unserer Grundvoraussetzungen ("arbeitserhaltend") ist verletzt, und zwar mit steigendem Verkehr immer mehr. Es gibt weitere ähnliche Effekte; insgesamt sind die Flaschenhalsüberlegungen nur bei Ausschluß von Blockierungen anwendbar.

LEERSEITE