Kapitel 3

Analytische Techniken für diskrete Systeme

In den bisherigen Beispielen spielten Warteschlangen und Warteschlangennetze eine zentrale Rolle, die meisten behandelten Beispiele waren von dieser Art. Das Verhalten wurde untersucht, indem der unterliegende stochastische Prozess nachgespielt wurde, d.h. es wurden mögliche Trajektorien beobachtet und aus ihnen mittels stochastischer Methoden Resultate geschätzt. Die Grenzen dieses Vorgehens wurden bereits dargestellt. So können Aussagen nur mit bestimmter Wahrscheinlichkeit, nicht aber mit Sicherheit gemacht werden. Weiterhin kann der Aufwand sehr hoch sein und gewisse Werte, wie sehr kleine Wahrscheinlichkeiten, können gar nicht per Simulation ermittelt werden. Insgesamt sollte immer die folgende Aussage gelten: *Simulation sollte immer das letzte Mittel sein, falls andere effizientere Techniken versagen oder nicht anwendbar sind.*

Damit stellt sich die Frage, welche anderen Techniken der Modellanalyse existieren? Von zentraler Bedeutung sind Methoden für einzelne Stationen und Warteschlangennetze. Diese Ansätze haben eine lange Tradition und gehen auf Arbeiten zur Dimensionierung von Telefonnetzen zu Beginn des vorherigen Jahrhunderts zurück. Ein Standardwerk über die Analyse von einfachen Wartesystemen ist [10]. Als Übersicht über Verfahren für Warteschlangennetze sei [13] empfohlen. Für den hier behandelten Stoffumfang reicht aber auch ein Blick in [3, Kap. 6] aus. Wir werden uns auf eine kurze Übersicht über die Thematik beschränken, weitere Details werden in der im Wintersemester angebotenen Vorlesungen "Kapazitätsplanung und Leistungsbewertung verteilter Systeme" behandelt.

Eine analytische Analyse ist in der Regel nur für relativ einfache Modelle und damit für eine eingeschränkte Modellklasse überhaupt möglich. Dies bedeutet, dass abstraktere Modelle entstehen müssen, die notwendigen Berechnungen aber auch deutlich effizienter als eine simulative Analyse sind. Darüber hinaus liefern sie exakte Resultate und keine Schätzer. Die Vorteile analytischer Modelle liegen damit auf der Hand. Der Aufwand der Modellbildung und Datenerhebung sinkt deutlich, da abstrakte Modelle eingesetzt werden. Gleichzeitig sinkt der Analyseaufwand. Dem gegenüber steht die eingeschränkte Modellierungsmächtigkeit. So lassen sich gewisse Phänomene nicht oder nur unzureichend mit analytischen Modellen abbilden. Inwieweit analytische Modelle für eine konkrete Problemstellung einsetzbar sind, ist immer eine Einzelfallentscheidung.

In diesem Kapitel werden zuerst Methoden zur Analyse einzelner Stationen vorgestellt und anschließend wird ein kurzer Einblick in die Analyse offener Warteschlangennetze gegeben.

3.1 Einfache Stationen

Einfache Stationen sind die Basis der Analyse von Warteschlangennetzen. Sie bilden eine Generalisierung unseres bereits mehrfach behandelten einfachen Schalters. Es gibt mehrere deutsche und englische Bezeichnungen für die Modelle und die bearbeiteten Aufträge/Jobs/Kunden. Wir wollen bei den Begriffen Station und Auftrag bleiben. Abbildung 3.1 zeigt die Struktur einer solchen

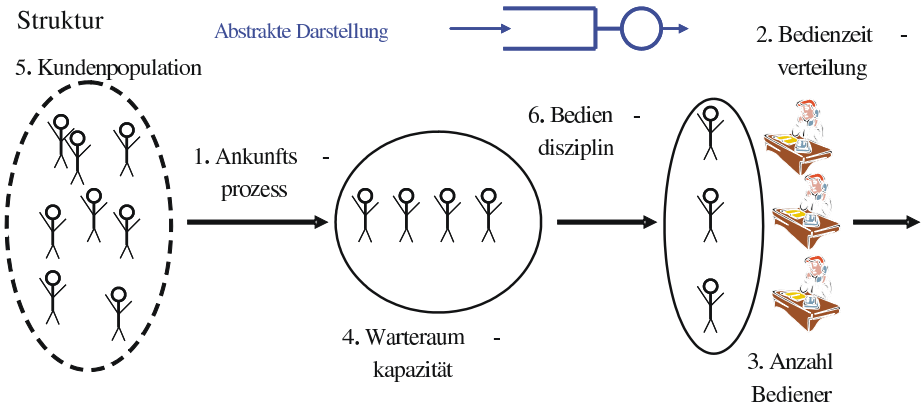

Abbildung 3.1: Struktur einer einfachen Station.

Station und klassifiziert die verschiedenen Teile. Üblicherweise wird eine Station in der einfachen abstrakten Darstellung gezeichnet und durch ein Sechstupel von Buchstaben und Zahlen beschrieben. Dieses Sechstupel bezeichnet man als Kendall-Notation. Bevor die Notation vorgestellt wird, sollen die in der Abbildung angegebenen sechs Elemente, die auch die Grundlage der Kendall-Notation sind, näher klassifiziert werden.

1. Der Ankunftsprozess beschreibt den stochastischen Prozess zur Generierung der Auftragsankünfte. Man unterscheidet zwischen stochastischen Prozessen und unabhängigen Ankünften, die durch eine Zwischenankunftszeitverteilung beschrieben sind. Wie betrachten nur den letzteren Fall. Für die Darstellung der üblichen Zwischenankunftszeitverteilungen werden Buchstaben verwendet. Folgende Verteilungen sind besonders verbreitet:

   - Exponential-Verteilung (M)
   - Erlang k-Verteilung (E_k)
   - Hyperexponential-Verteilung mit k Phasen (H_k)
   - Deterministische-Verteilung (D)
   - Allgemeine Verteilung (G) oder (GI), falls zusätzlich die Unabhängigkeit der Ankünfte herausgestellt werden soll.

   (a) Die bisherigen Verteilungen beschreiben einzelne Ankünfte. Man kann die Notation erweitern, um Gruppen-Ankünfte zur beschreiben. Die symbolische Darstellung lautet dann A^B, wobei $A \in \{M, E_k, H_k, D, G, GI, \ldots\}$ die Zwischenankunftszeit spezifiziert und $B \in \{M, Geo, D, G, GI, \ldots\}$ die diskrete Gruppengröße beschreibt. Im letzteren Fall steht M für einen Poisson-Prozess, Geo für eine geometrische Verteilung und G für eine allgemeine diskrete Verteilung. M^M bedeutet als die Ankunft von Poisson-verteilten Gruppengrößen mit exponentiell verteilten Zwischenankunftszeiten.

2. Die zweite Komponente beschreibt die Bedienzeitverteilung. Es wird in der Regel von unabhängigen Bedienzeiten ausgegangen, die mit den selben Buchstaben wie die Zwischenankunftszeiten charakterisiert werden. Auch bei den Bedienzeiten können Gruppen-Bedienung ähnlich zu Gruppenankünften definiert werden. In diesem Fall beginnt die Bedienung erst, wenn eine genügende Zahl von Aufträgen vorhanden ist und alle Aufträge werden auf einmal bedient.

3. Die dritte Komponente spezifiziert die Anzahl der vorhandenen Bediener. Es wird davon ausgegangen, dass alle Bediener identisch und damit ununterscheidbar sind. Wird eine unendliche Zahl spezifiziert, so bedeutet dies, dass für jeden Auftrag immer eine Bediener vorhanden ist. Man spricht in diesem Fall auch von einer Verzögerungsstation.

4. Die Größe des Warteraums wird durch die vierte Komponente beschrieben. Es wird davon ausgegangen, dass alle Aufträge bis zum Ende ihrer Bedienung im Warteraum bleiben. Es gilt damit Größe Warteraum $\leq$ Anzahl Bediener. Aufträge, die eintreffen, wenn der Warteraum vollständig gefüllt ist, werden abgewiesen und gehen damit in der Regel verloren. Die folgenden Sonderfälle werden unterschieden:

   - Falls Anzahl Bediener = Größe Warteraum spricht man von einem Verlustsystem
   - Falls Anzahl Bediener < Größe Warteraum $< \infty$ spricht man von einem Warte-/Verlustsystem.

5. Die fünfte Komponente spezifiziert die Kundenpopulation im System. Alle Kunden können potenziell beim Bediener eintreffen. Neben der maximalen Population wird durch die Population die Ankunftsintensität bestimmt. Bei einer unendlichen Systempopulation geht man davon aus, dass die Zwischenankunftszeit die Zeit zwischen zwei Ankünften an der Station beschreibt. Bei einer endlichen Population beschreibt die Zwischenankunftszeit die Zeit, die für einen Auftrag zwischen dem Verlasen der Station und der nächsten Ankunft vergeht. Wenn sich also K Aufträge im System befinden und die Ankünfte sind vom Typ M, so laufen bei Population 0 in der Station K Exponential-Verteilungen parallel und die nächste Ankunft erfolgt nach dem Minimum der ermittelten Zeiten. Befinden sich dagegen 2 Aufträge in der Station so laufen nur $K-2$ Exponential-Verteilungen parallel, um die nächste Ankunft zu bestimmen. Damit hängt die Ankunftsintensität von der Population in der Station ab.

6. Die letzte Komponente beschreibt die Bediendisziplin, nach der Aufträge aus dem Warteraum ausgewählt werden, um in die Bedienung zu gelangen. Es gibt auch hier wieder eine Vielzahl von Disziplinen, die durch die jeweiligen Abkürzungen gekennzeichnet sind.

   - FCFS: Die Aufträge werden in der Reihenfolge ihres Eintreffens abgearbeitet.
   - Random: Es wird zufällig ein Auftrag für die Bedienung ausgewählt.
   - LCFS (last come first served): Der zuletzt angekommene Auftrag wird als erster bedient.
   - PS (processor sharing): Die Bedienkapazität wird unter allen warteneden Aufträgen gleichmäßig aufgeteilt. Wenn also ein Bediener k Aufträge gleichzeitig bedient, so verlängert sich die Bedienzeit jedes Auftrags um den Faktor k. Verlässt während der Bedienung eine Auftrag die Station, so verringert sich die Restzeit der übrigen Aufträge um den Faktor $(k-1)/k$. Analog verlängert sich die Restzeit um den Faktor $(k+1)/k$, wenn ein neuer Auftrag während der Bedienung eintrifft.

Neben diesen explizit formulierten Verhaltensbeschreibungen gibt es noch implizite Annahmen. So wird angenommen, dass die Bediener immer arbeiten wenn Aufträge vorhanden sind, aber ein Auftrag auch nur von einem Bediener bedient werden kann. Da die Bediener ununterscheidbar sind, wird ein Auftrag zufällig von einem Bediener bedient. Die Beschreibung bedingt, dass die Bedienstrategie PS nur bei einem Bediener definiert ist.

Die Kendall-Notation baut auf den auf den obigen 6 Charakteristika auf und ist von der Form $A/B/c/N/K/SD$ mit

- A Zwischankunftszeit nach obiger Notation
- B Bedienzeit nach obiger Notation.
- c Anzahl Bediener als ganze Zahl.
- N Kapazität des Warteraums als ganze Zahl mit Voreinstellung $N = \infty$.
- K Gesamtpopulation als ganze Zahl mit Voreinstellung $K = \infty$.
- SD Bedienstrategie nach den obigen Abkürzungen mit Voreinstellung $FCFS$.

In den Fällen, in denen Voreinstellungen vorhanden sind, können die zugehörigen Komponenten weggelassen werden. So steht $M/M/1$ für $M/M/1/\infty/\infty/FCFS$ und beschreibt eine Station mit exponentiell verteilten Zwischenankunftszeiten, exponentiell verteilten Bedienzeiten, einem Bediener, einen undendlichen Warteraum, einer unendlichen Population im System und einer Bedienung nach Ankunft der Aufträge. $M/M/1/K$ unterscheidet sich vom vorherigen System dadurch, dass der Warteraum eine Kapazität von K und nicht von ∞ hat.

Stationen können für sehr unterschiedliche Systeme zur Modellierung verwendet werden. Die folgende Tabelle zeigt einige mögliche Interpretationen:

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

Auf der Analyseebene spielt die Interpretation des Systems keine Rolle. Die Analyse kann wieder über einen endlichen Zeitraum erfolgen (man spricht von einer transienten Analyse) oder über einen unendlichen Zeitraum (man spricht von einer stationären Analyse). Meistens wird die stationäre Analyse betrachtet, da nur für diesen Fall analytische Berechnungen möglich sind. Wir beschränken uns deshalb auf die stationäre Analyse und setzen dabei implizit voraus, dass das untersuchte Modell auch einen stationären Zustand erreicht. Die dazu notwendigen Bedingungen ergeben sich aus den einzelnen Berechnungen. Im Gegensatz zur Simulation, bei der stationäre Resultate aus einer endlichen Beobachtung abgeleitet werden müssen, können im analytischen Fall exakte Ergebnisse für eine, nicht real durchführbare, Beobachtung über einen unendlichen Zeitraum bestimmt werden, sofern das Modell gewissen Einschränkungen unterliegt.

Im Folgenden werden zuerst einige allgemeine Notationen eingeführt und Resultate hergeleitet. Anschließend wird die Analyse spezieller einfacher Systeme beschrieben.

Allgemeinen Notationen und Resultate

Wir betrachten eine Station mit unabhängigen Zwischenankunfts- und Bedienzeiten mit Erwartungswerten $E(A)$ und $E(B)$. Die Ankunfts- und Bedienrate lauten dann $\lambda = E(A)^{-1}$ und $\mu = E(B)^{-1}$. Da Ankunfts- und Bedienzeiten in der Regel Zufallsvariablen sind, zeigt das System ein stochastisches Verhalten und mögliche Resultatgrößen sind ebenfalls Zufallsvariablen. Wir betrachten die folgenden Resultatgrößen:

- X der Durchsatz des Systems beschreibt die Anzahl Aufträge, die pro Zeiteinheit die Station verlassen.
- Q ist die Auftragspopulation in der Station.
- R ist die Verweilzeit eines Auftrags in der Station.
- U ist die Auslastung des oder der Bediener, d.h. der Zeitanteil, den ein Bediener arbeitet. Bei mehreren Bedienern wird angenommen, dass alle Bediener sich die anfallende Arbeit zu gleichen Anteilen teilen und damit eine identische Auslastung haben.

Wie schon bei der Simulation werden nur die Erwartungswerte der obigen Größen ermittelt. Zusätzlich wird noch $p(i)$, die Wahrscheinlichkeit, dass sich i Aufträge an der Station befinden bestimmt. Mit der Hilfe von $p(i)$ wird natürlich auch die Verteilung der Population Q ermittelt.

Wir betrachten zuerst einige Zusammenhänge, die sehr allgemein gelten. Dies bedeutet insbesondere, dass keine weiteren Annahmen über die Struktur des System gemacht werden.

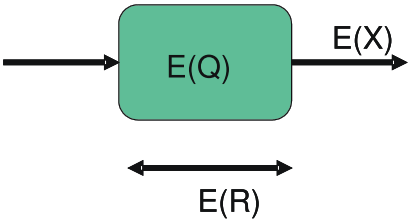

Abbildung 3.2: Allgemeine Darstellung des Gesetzes von Little.

In Systemen mit einem Bediener gilt

$$E(U) = 1 - p(0) \ ,$$

da der Bediener immer dann arbeitet, wenn mindestens ein Auftrag wartet. Der Erwartungswert der Population ergibt sich bei Kenntnis der $p(i)$ als

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

Der Erwartungswert des Durchsatzes ergibt sich aus

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

da ein belegter Bediener mit Rate μ Aufträge bedient. Wenn sich weniger als m Aufträge an der Station befinden, so hat jeder seinen eigenen Bediener, bei m oder mehr Aufträgen sind alle Bediener belegt.

Das nun folgende Gesetz von Little ist eines der fundamentalen Gesetzte der Analyse von Warteschlangennetzen, da es unter sehr allgemeinen Bedingungen gilt. Zur Erläuterung betrachten wir ein System, in dem Aufträge bedient werden. Die interne Struktur bleibt dem Betrachter verborgen (black box). Sei Q die Zahl der Aufträge im System und $E(Q)$ der Erwartungswert. R sei die Verweilzeit der Aufträge mit Erwartungswert $E(R)$ und X sei der Durchsatz mit Erwartungswert $E(X)$. Wir nehmen an, dass keine Aufträge im System verloren gehen (d.h. das System nicht mehr verlassen) und alle Erwartungswerte endlich sind. Unter diesen Bedingungen gilt der folgende Zusammenhang zwischen den drei Erwartungswerten.

$$E(Q) = E(X) \cdot E(R) \tag{3.1}$$

Der Zusammenhang gilt unabhängig von den Verteilungen und wird als Gesetzt von Little bezeichnet. Bevor wir den Beweis etwas näher betrachten, soll kurz die Bedeutung der Gleichung erläutert werden. Bei vielen realen Systemen kann man einzelne Größen beobachten und andere nicht. So ist es zum Beispiel oft recht einfach die mittlere Population und den Durchsatz zu messen, während eine Messung der Verweilzeit den Zugriff auf individuelle Daten jedes einzelnen Auftrags erfordert. Durch Umstellen der obigen Formel kann aus $E(X)$ und $E(Q)$ die mittlere Verweilzeit berechnet werden. Auch bei der Analyse können $E(X)$ und $E(Q)$ bei Kenntnis von $p(i)$ berechnet werden und $E(Q)$ kann aus den beiden Werten ermittelt werden.

Betrachten wir als Beispiel unseren Fachbereich. Die Zahl der Studierenden liegt bei ungefähr 2500, wenn wir eine Anfängerzahl von 400 pro Jahr voraussetzen, so wäre die mittlere Verweilzeit ungefähr 6.25 Jahre. Auch wenn dieser Werte nicht so weit von der mittleren Studeindauer entfernt ist, messen wir hier nicht die Studiendauer, sondern die Verweilzeiten inklusive der Studienabbrecher. Zur Ermittlung der Studiendauer müsste man die Zahl der Diplomanden als Durchsatz wählen und nur die Studierenden zählen, die ihr Studium später abschließen. Dieser Wert ist aber nicht der Statistik zu entnehmen.

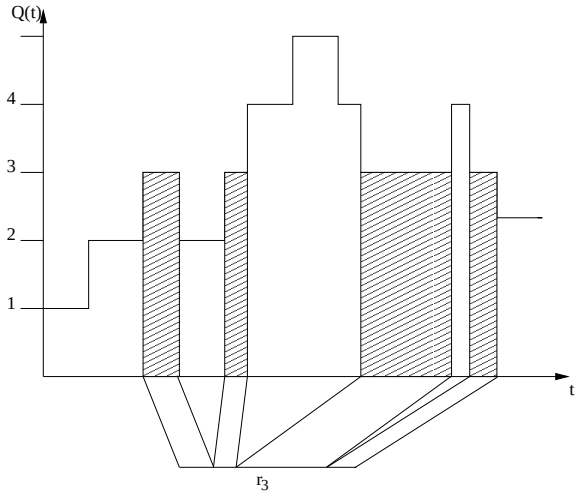

Abbildung 3.3: Trajektorie des Systems zum Beweis des Satzes von Little.

Es soll nun ein Beweis für den Satz von Little kurz gezeigt werden. Es gibt viele unterschiedliche Beweise. Wir verwenden einen anschaulichen Ansatz, der auf der Beobachtung des Systems über einem endlichen Zeitraum beruht. Dazu betrachten wir das System im Intervall $[0, T]$. T soll dabei so groß gewählt werden, dass die Anzahl der Aufträge, die das System betreten haben und die Anzahl der Aufträge, die das System verlassen haben, sich prozentual kaum unterschieden.

Sei f_k der prozentuale Anteil der Beobachtungszeit, zu dem sich k Aufträge im System befanden. Unter den gemachten Annahmen gilt $\lim_{T\to\infty} f_k = p(k)$. Weiterhin sei $r_k = f_k \cdot T$, die Zeit zu der sich k Aufträge während der Beobachtung im System befanden. Seien $\bar{Q}$ und $\bar{X}$ die mittlere Population und der mittlere Durchsatz im Beobachtungsintervall. Auch hier gilt $\lim_{T\to\infty} \bar{Q} = E(Q)$ und $\lim_{T\to\infty} \bar{X} = E(X)$. C_0 sei die Gesamtzahl der Aufträge, die das System im Beobachtungsintervall verlassen hat. Dann gilt

$$\bar{Q} = \sum_{k=1}^{\infty} k \cdot f_k = \sum_{k=1}^{\infty} k \cdot r_k/T \quad \text{und} \quad \bar{X} = C_0/T \ .$$

Wenn wir die vorherige Darstellung von $\bar{Q}$ mit C_0/C_0 multiplizieren, erhalten wir

$$\bar{Q} = \frac{C_0}{T} \cdot \frac{\sum_{k=1}^{\infty} k \cdot r_k}{C_0} = \bar{X} \cdot \bar{R}$$

Die letzte Umformung gilt, da $\sum_{k=1}^{\infty} k \cdot r_k$ die gesamte kumulierte Verweilzeit von Aufträgen im System beschreibt, Division durch die Anzahl der Aufträge liefert die Verweilzeit eines Auftrags. Da die Zusammenhänge für alle Mittelwerte gelten und die Mittelwerte gegen die Erwartungswerte konvergieren, gilt der Zusammenhang auch für die Erwartungswerte.

Das $M/M/1$-System

Nachdem bisher allgemeine Gesetzte hergeleitet wurden, soll nun ein spezielles Modell vollständig analysiert werden, um die konkreten Werte zu berechnen. Die einfachste Variante ist das $M/M/1$-System. Die Zwischenankunftszeiten sind exponentiell verteilt mit Verteilungsfunktion $FA(t) = 1 - e^{-\lambda t}$. Damit bilden die Ankünfte einen Poisson-Prozess, so dass im Intervall $[0, t)$ mit Wahrscheinlichkeit $P[Y = k] = e^{-\lambda t}(\lambda t)^k/k!$ genau k Ankünfte beobachtet werden können. Die Bedienzeiten sind ebenfalls exponentiell verteilt mit Verteilungsfunktion $FB(t) = 1 - e^{-\mu t}$. Die Station hat einen Bediener, einen Warteraum unendlicher Kapazität und die potenzielle Population ist unendlich. Über die Bedienstrategie machen wir erst einmal keine Angaben, wenn es nötig ist sei die Bedienstrategie als FCFS festgelegt.

max	$t = 100$				$t = 1000$				$t = 2000$		
λ	min	max	$\hat{Q}$	$\hat{S}$	min	max	$\hat{Q}$	$\hat{S}$	max	$\hat{Q}$	$\hat{S}$
0.05	0	2	0.100	0.333	0	2	0.070	0.293			
0.5	0	8	1.060	1.693	0	8	0.970	1.410			
0.95	0	31	7.510	6.711	0	67	18.07	16.89	84	21.71	20.47
0.99	0	43	10.16	8.178	0	116	31.70	26.51	177	41.16	37.43

Tabelle 3.1: Ergebnisse der Simualtionen für $t = 100$, $t = 1000$ und $t = 2000$.

Das Modell entspricht genau unserem einfachen Schalter aus Abschnitt 2.1. Damit ist eine Simulation einfach realisierbar. Zur Ermittlung stationärer Resultate können entweder mehrere Replikationen durchgeführt werden oder ein langer Simulationslauf ausgewertet werden. Auch wenn in den meisten Fällen ein langer Simulationslauf sinnvoller ist (siehe auch Abschnitt 2.5), soll erst einmal das Beispiel für mehrere Replikationen betrachtet werden. Sei dazu die Bedienrate $\mu = 1.0$ fest gewählt. Ziel der Analyse ist die Ermittlung von $E(Q)$. Wir simulieren dazu das Modell bis zu einem Zeitpunkt t und ermitteln dann den Wert von Q. Diese Simulation wird 100mal wiederholt und aus den Ergebnissen der Mittelwert- und Varianzschätzwert ermittelt. Tabelle 3.1 zeigt die Ergebnisse der Simulationsläufe. Für die kleineren Werte von λ scheint ein Wert von $t = 1000$ ausreichen zu sein, während bei den beiden großen Werten von λ nicht einmal klar ist, ob $t = 2000$ ausreicht. Die Konfidenzintervalle für $\lambda = 0.05$ lautet $[0.022, 0.118]$ und für $\lambda = 0.5$ $[0.737, 1.135]$, wenn jeweils die Daten zum Zeitpunkt $t = 1000$ erhoben wurden und die Quantile der Normalverteilung für die Berechnung verwendet werden. Für $\lambda = 0.95$ und $\lambda = 0.99$ lauten die Konfidenzintervalle $[18.34, 25.10]$ bzw. $[40.98, 53.34]$, wenn die Daten zum Zeitpunkt $t = 2000$ verwendet werden. Das Beispiel zeigt, dass wir auch für relativ große Beobachtungszahlen breite Konfidenzintervalle bekommen. Um das Konfidenzintervall für $\lambda = 0.5$ auf eine Breite von 10% des ermittelten Mittelwertschätzers zu reduzieren werden ca. 1800 Replikationen benötigt. Für die größeren Werte von λ werden entsprechend mehr Replikationen benötigt.

Auch die Beobachtung eines langen Simulationslaufs ändert nichts grundsätzlich an den Problemen. Im Gegenteil, bei einer hohen Auslastung (d.h. großen Werten von λ) sind die Werte stärker korreliert und es müssen große Batches gebildet werden, so dass sich neben der Erkennung des Endes der transienten Phase auch das Problem der Erkennung einer ausreichenden Batch-Größe ergibt. Immer dann, wenn hoch ausgelastete Systeme oder selten auftretende Ereignisse analysiert werden sollen, ergeben sich bei der Simulation Probleme.

Diese Problematik soll an einem weiteren Beispiel etwas formaler analysiert werden. Wir betrachten dazu die Simulation eines technischen Systems z.B. einer Sonde, die zu einem anderen Planeten geschickt wird, über ein Intervall $[0, T]$. Es soll analysiert werden, ob das System zum Zeitpunkt T noch funktionsfähig ist. Am Beispiel der Sonde würde dies bedeuten, dass die technischen Geräte bei Ankunft auf dem Planeten noch arbeiten. Zur Analyse wird das Modelle des Systems wiederholt bis zum Zeitpunkt T simuliert. Sei y_i das Ergebnis der $i$ten Replikation und Y_i die Zufallsvariable, die das Resultat beschreibt. Es gelte $y_i = 1$, falls die Simulation am Ende der $i$ten Replikation ein defektes System liefert und 0 sonst. Zu ermitteln ist nun $P[Y = 1] = p$. Für vernünftig entworfene Systeme sollte $p \ll 1$ gelten. Ziel der Simulation ist die Bestimmung von p mit einer Genauigkeit von $\pm 10\%$ des ermittelten Wertes und einer Signifikanzwahrscheinlichkeit von 99%. Wie viele Abläufe müssen dafür simuliert werden?

Der Resultatschätzer $\tilde{p}$ lautet

$$\tilde{p} = \frac{1}{n} \cdot \sum_{i=1}^{n} Y_i \text{ und } \hat{p} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$$

lautet der Schätzwert. Es gilt $E(\tilde{p}) = p$ und $\sigma^2(\tilde{p}) = p \cdot (1-p)/n$. Damit ergibt sich ein Konfidenzintervall von

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

Da $E(\tilde{S}^2) = n \cdot \sigma^2(\tilde{p}) = p \cdot (1-p)$ entspricht der Erwartungswert der halben Breite des Konfidenzintervalls.

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

Damit soll gelten

$$2.576 \cdot \sqrt{p \cdot (1-p)/n} \leq 0.1 \cdot \hat{p} \ .$$

Da $\lim_{n\to\infty} \hat{p} = p$ gilt auch

$$n \approx 100 \cdot 2.576^2 \cdot (1-p)/p = 663.58 \cdot (1-p)/p \ .$$

Je kleiner p ist, desto größer muss n gewählt werden, um das gewünschte Konfidenzintervall zu ermitteln. Die folgenden Beispiele zeigen dies eindrucksvoll.

$$\begin{array}{ll} p = 0.1 \ \Rightarrow \ n \approx 5973 & p = 0.001 \ \Rightarrow \ n \approx 662917 \\ p = 10^{-6} \ \Rightarrow \ n \approx 6.63 \cdot 10^8 & p = 10^{-8} \ \Rightarrow \ n \approx 6.63 \cdot 10^{10} \end{array}$$

Selbst auf schnellen Rechnernen und mit parallelen Berechnungen lassen sich nicht mehr als 60 Milliarden Replikationen simulieren, auch wenn eine einzelne Replikation sehr schnell simuliert werden kann. Bei komplexeren Modellen erfordert aber auch schon die Simulation einer einzigen Replikation mehrere Minuten. Damit bleibt die Frage, inwieweit die Bestimmung solch kleiner Wahrscheinlichkeiten von Interesse ist. In heutigen technischen Systemen sind Wahrscheinlichkeiten der Ordnung 10^{-8} und auch 10^{-9} durchaus üblich. So liegt die geforderte Verfügbarkeit technischer Systeme heute bei 10^{-6} und im sicherheitskritischen Bereich sogar bei 10^{-8}. In Rechnernetzen sollen Verlustwahrscheinlichkeiten bei kritischen Diensten im Bereich von 10^{-9} liegen. Diese Beispiele zeigen die Grenzen der Simulation und damit die Möglichkeiten analytischer Modelle.

Die analytische Berechnung wird nun am Beispiel des bereits beschriebenen und simulierten $M/M/1$-Systems vorgeführt. Dazu betrachten wir zuerst einige Eigenschaften der Exponentialverteilung (siehe auch Seite 64). Wie gezeigt ist die Exponentialverteilung die einzige kontinuierliche Verteilung, die gedächtnislos ist, d.h. $F(t+x|x) = F(t)$ für alle $t, x \geq 0$ und exponentiell-verteilte Zufallsvariablen. Darüber hinaus gilt das Prinzip der Additivität. Das Minimum zweier Exponentialverteilungen mit Raten μ und λ ist wieder exponentiell-verteilt mit Rate $\mu + \lambda$. Wenn eine Ereignis beobachtet wird, so ist dies mit Wahrscheinlichkeit $\mu/(\mu+\lambda)$ aus der ersten Verteilung mit Rate μ und mit Wahrscheinlichkeit $\lambda/(\mu+\lambda)$ aus der zweiten Verteilung mit Rate λ.Das Prinzip lässt sich natürlich auf eine beliebige Anzahl von Exponentialverteilungen erweitern.

Wir können uns kurz über die Bedeutung der beiden Eigenschaften Gedanken machen. Die Gedächtnislosigkeit bedingt, dass die Wartezeit auf ein Ereignis, welches nach einer exponentiell-verteilten Zeit eintritt, ausgehend von einem beliebigen Zeitpunkt bevor das Ereignis eintritt, unabhängig von der bereits gewarteten Zeit ist. Additivität bedeutet, dass beim Warten auf ein Ereignis aus einer Menge von Ereignissen mit exponentiell verteilter Eintrittszeit eine exponentiell verteilte Zeit gewartet werden muss, deren Rate sich aus der Summe der Raten der einzelnen Verteilungen ergibt. Beide Beobachtungen spielen auch bei der Analyse des $M/M/1$-Systems eine zentrale Rolle.

Das $M/M/1$-System beschreibt eine spezielle Form eines Markov-Prozesses, nämlich einen so genannten Geburts-Todesprozess. Dazu betrachten wir den Zustandsraum des Systems $S = \{0, 1, 2, \ldots\}$, wobei $i \in S$ die Station mit i Aufträgen beschreibt. Der Zustand ändert sich durch Ereignis, die im Kontext der Markov-Prozesse als Transitionen bezeichnet werden. Es gibt zwei Typen von Transitionen, Ankünfte und Abgänge (oder Bedienenden). Eine Ankunft ist immer möglich und ändert den Zustand von i nach $i+1$. Ein Abgang ist nur in Zuständen $i > 0$ möglich und ändert den Zustand von i nach $i-1$.

Die Prinzipien der Analyse von Markov-Prozessen sollen hier nur an den konkreten Beispielen einzelner Stationen beschrieben werden. Das allgemeine Konzept geht deutlich darüber hinaus. Eine gute Übersicht über die Analyse von Markov-Prozessen liefert [17]. Markov-Prozesse können als markierte Graphen dargestellt werden. Die Zustände bilden die Knoten und die Transitionen die Kanten. Kanten werden zusätzlich mit den Transitionsraten, d.h. den Raten der zugehörigen

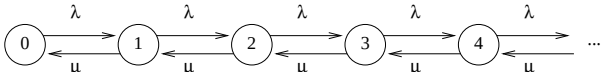

Abbildung 3.4: Markov-Prozess für ein $M/M/1$-System.

Exponentialverteilungen, gewichtet. Abbildung 3.4 zeigt die graphische Darstellung des Markov-Prozesses, der durch ein $M/M/1$-System beschrieben wird.

Da Bedienzeiten und Zwischenankunftszeiten exponentiell-verteilt sind, ergibt sich folgendes Verhalten in den einzelnen Zuständen:

- Im Zustand $i = 0$ vergeht eine exponentiell verteilte Zeit mit Rate λ bis zur Ankunft des nächsten Auftrags und des damit verbundenen Übergangs in Zustand 1.
- Im Zustand $i > 0$ vergeht eine exponentiell verteilte Zeit mit Rate $\mu + \lambda$ bis zur nächsten Zustandsänderung. Mit Wahrscheinlichkeit $\mu/(\mu + \lambda)$ ist die nächste Zustandsänderung ein Abgang mit Nachfolgezustand $i - 1$ und mit Wahrscheinlichkeit $\lambda/(\mu + \lambda)$ eine Ankunft mit Nachfolgezustand $i + 1$.

Auf Grund der Gedächtnislosigkeitseigenschaft kann das beschriebene Verhalten immer beobachtet werden, wenn das Modell in einem Zustand ist, unabhängig davon, wie lange das Modell bereits in diesem Zustand war. Anschaulich verringert sich die Wahrscheinlichkeit in einem Zustand durch den Abfluss über ausgehende Kanten. Der zugehörige Wahrscheinlichkeitsfluss entspricht dem Produkt aus Wahrscheinlichkeit im Zustand zu sein und den Raten an den abgehenden Kanten. Gleichzeitig gibt es eine Zufluss durch eingehende Kanten. Auch hier gilt wieder, dass der Fluss der Wahrscheinlichkeit des Quellzustandes multipliziert mit der Rate der Kante entspricht. Da wir am stationären Gleichgewicht des Modells interessiert sind, ist die Verteilung zu ermitteln, bei der sich die Zustandswahrscheinlichkeiten nicht mehr ändern. Dies bedeutet, dass der Abfluss und der Zufluss an Wahrscheinlichkeit in jedem Zustand identisch sein muss. Auf Grund der speziellen Struktur des Markov-Prozesses für $M/M/1$-Systeme reicht es aus benachbarte Zustände zu betrachten. Für Zustand 0 gilt $p(0) \cdot \lambda = p(1) \cdot \mu$ und allgemein gilt $i \geq 1$

$$p(i) \cdot \mu = p(i-1) \cdot \lambda \;\Rightarrow\; p(i) = p(i-1) \cdot \left(\frac{\lambda}{\mu}\right) \;\Rightarrow\; p(i) = p(0) \cdot \left(\frac{\lambda}{\mu}\right)^i \; . \tag{3.2}$$

Mit der Abkürzung $\rho = \lambda/\mu$ gilt $p(i) = p(0) \cdot \rho^i$. Falls wir also $p(0)$ kennen würden, so könnten alle anderen Wahrscheinlichkeiten einfach hergeleitet werden und aus diesen die Leistungsgrößen ermittelt werden, wie bereits gezeigt wurde.

Zur Bestimmung von $p(0)$ hilft die folgende Überlegung. Die $p(i)$ beschreiben eine Wahrscheinlichkeitsverteilung, d.h. die Summe über alle Werte muss 1 ergeben. Damit gilt

$$\sum_{i=0}^{\infty} p(i) = p(0) \cdot \sum_{i=0}^{\infty} \rho^i = 1 \;\Rightarrow\; p(0) = \frac{1}{\sum_{i=0}^{\infty} \rho^i} \; .$$

Die einzelnen Wahrscheinlichkeiten ungleich 0 und damit berechenbar, wenn die Reihensumme endlich ist. Da es sich um eine geometrische Reihe handelt gilt

$$\sum_{i=0}^{\infty} \rho^i = (1-\rho)^{-1} \;\text{ falls }\; \rho < 1 \; .$$

Damit resultiert aus der Herleitung der stationären Wahrscheinlichkeitsverteilung die Bedingung $\rho < 1$. Diese soll kurz auf Modellebene interpretiert werden.

- Falls $\lambda > \mu$ ist, so kommen im Mittel mehr Aufträge an als abgearbeitet werden können. Für einen unendlichen Zeithorizont wächst die unerledigte Arbeit und damit die Auftragszahl in der Warteschlange über alle Grenzen und ein stationärer Zustand existiert nicht.

- Für $\lambda = \mu$ zeigt die mathematische Herleitung genau das selbe Verhalten. Dies kann man dadurch erklären, dass wir es mit einem System mit stochastischen Schwankungen zu tun haben. Im Mittel kommt zwar genauso viel Arbeit hinzu, wie abgearbeitet werden kann, auf Grund der stochastischen Schwankungen wird es immer mal wieder Perioden geben, in denen die Station leer ist und Bedienkapazität verloren geht. Die so verloren gegangene Bedienkapazität kann nicht in späteren Phasen aufgeholt werden, da die Last der Bedienkapazität entspricht. Über einen unendlichen Zeitraum betrachtet führt dies dazu, dass keine stationärer Zustand existiert.

Damit kann eine stationäre Analyse nur für den Fall $\lambda < \mu$ bzw. $\rho < 1$ durchgeführt werden. Für diesen Fall sollen nun die einzelnen Leistungsgrößen berechnet werden. Die bisher vorgestellten allgemeinen Formeln basieren zum Teil auf unendlichen Summen, die nicht vollständig ausgewertet können. Für den speziellen Fall des $M/M/1$-Systems lassen sich aber geschlossene Berechnungen herleiten. Für die Auslastung gilt

$$U = 1 - p(0) = \rho \ .$$

der Durchsatz entspricht gerade der Ankunftsrate $E(X) = \lambda$, da keine Aufträge verloren gehen. Die Herleitung der mittleren Population ist etwas komplexer.

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

Die mittlere Verweilzeit im System lässt sich über den Satz von Little bestimmen.

$$E(R) = \frac{E(Q)}{\lambda} = \frac{1}{\mu - \lambda}$$

Auf dieser Basis lassen sich weitere Leistungsgrößen berechnen. Da die Wahrscheinlichkeiten $p(i)$ bekannt sind, kann die Varianz der Population in der Station berechnet werden. Allgemein gilt

$$\sigma^2(Q) = E(Q^2) - (E(Q))^2$$

und mit

$$\begin{aligned} E(Q^2) &= (1-\rho) \cdot \sum_{i=0}^{\infty} i^2 \cdot \rho^i && = (1-\rho) \cdot \rho \cdot \tfrac{d}{d\rho} \sum_{i=0}^{\infty} i \cdot \rho^i \\ &= (1-\rho) \cdot \rho \cdot \tfrac{d}{d\rho} \tfrac{\rho}{(1-\rho)^2} && = (1-\rho) \cdot \rho \cdot \tfrac{(1-\rho)^2 + 2\rho \cdot (1-\rho)}{(1-\rho)^4} \\ &= \rho \cdot \tfrac{(1-\rho)+2\rho}{(1-\rho)^2} && = \tfrac{\rho + \rho^2}{(1-\rho)^2} \end{aligned}$$

folgt

$$\sigma^2(Q) = \frac{\rho + \rho^2}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} = \frac{\rho}{(1-\rho)^2} \ .$$

Wenn wir voraussetzen, dass immer nur ein Auftrag bedient wird, die Analyse auf wartende Aufträge beschränkt wird und Q_W die mittlere Anzahl wartender Aufträge ist, so gilt

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

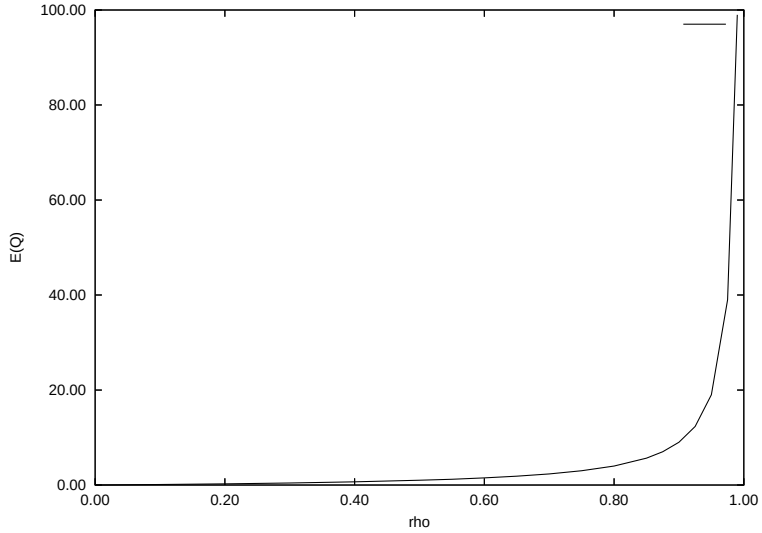

Abbildung 3.5: Verlauf von $E(Q)$ in Abhängigkeit von ρ für ein $M/M/1$-System.

Mit Hilfe des Satzes von Little folgt wieder die Wartezeit als

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

da der Durchsatz durch den Warteraum ebenfalls λ ist.

Es sollte beachtet werden, dass bisher keine Angaben über die Bedienreihenfolge in die Analyse eingeflossen sind. Dies bedeutet, dass die Ergebnisse für alle Bedienreihenfolgen gelten. Für die Strategie PS, bei der quasi alle Aufträge gleichzeitig bedient werden, können wir keine Aussagen über den Warteraum machen, die Aussagen bzgl. der mittleren Population und Verweilzeit in der Station gelten aber. Es wurden Ergebnisse für die Verteilung der Aufträge im System hergeleitet, aber keine Aussagen über Charakteristika der Verweilzeit gemacht, die über den Erwartungswert hinausgehen. Wenn solche Aussagen gemacht werden, so muss die Bedienstrategie berücksichtigt werden.

Der Verlauf von $E(Q)$wird in Abbildung 3.5 für variierendes ρ dargestellt. Für niedrige Auslastungen ist der Verlauf fast linear, während für $\rho > 0.8$ ein sehr schneller Anstieg beobachtet werden kann. Dieses Verhalten ist typisch für Stationen. Der prinzipiell Verlauf ist unabhängig von den konkreten Verteilungen, diese bestimmen nur den kritischen Punkt an dem der steile Anstieg beginnt.

Weitere $M/M/\ldots$-Systeme

Grundsätzlich lässt sich das für das $M/M/1$-System vorgestellte Vorgehen auch auf andere Systeme mit exponentiell verteilten Zwischenankunfts- und Bedienzeiten übertragen. Das prinzipielle Vorgehen ist dabei immer sehr ähnlich. Der Zustandsraum entspricht den natürlichen Zahlen oder einer endlichen Teilmenge bei Systemen mit endlicher Kapazität. Die Transitionen beschreiben immer einen Geburts-/Todesprozess. Es treten die folgenden Unterschiede zum $M/M/1$-System auf.

1. Die Bedienraten und Ankunftsraten können von der Population abhängen, indem
   (a) in Mehrbedienersystemen, die Bedienrate linear mit der Population steigt, bis alle Bediener belegt sind
   (b) durch eine endliche Population im System die Ankunftsrate fällt, wenn die Population in der Station steigt.

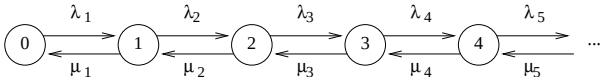

Abbildung 3.6: Allgemeiner Geburts-/Todesprozess.

2. Die Population der Station kann endlich sein.

3. Dazu können noch Kombinationen aus 2. und 3. auftreten.

Die allgemeinste Form eines Geburts-/Todesprozesses für eine Station mit unendlicher Kapazität und lastabhängigen Raten ist in Abbildung 3.6 zu sehen. Analog zum einfachen Fall mit konstanten Raten muss für die stationären Zustandswahrscheinlichkeiten

$$p(i) \cdot \mu_i = p(i-1) \cdot \lambda_{i-1} \;\Rightarrow\; p(i) = p(i-1) \cdot \left(\frac{\lambda_{i-1}}{\mu_i}\right) \;\Rightarrow\; p(i) = p(0) \cdot \prod_{j=1}^{i} \left(\frac{\lambda_{j-1}}{\mu_j}\right)$$

für $i > 0$ gelten. Die Wahrscheinlichkeit $p(0)$ resultiert wieder aus der Beobachtung, dass die Summe der Wahrscheinlichkeiten 1 ergibt, so dass

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

gilt. Um die unendliche Summe zu berechnen, müssen weitere Annahmen getroffen werden. Übliche Annahmen sind

- eine endliche Anzahl lastabhängiger Raten, so dass $\lambda_{i-1}/\mu_i = \lambda^*/\mu^*$ für $i > c$ (z.B. $M/M/c$-System)
- ein fallender Quotient $\lambda_{i-1}/\mu_i < \lambda_i/\mu_{i+1} < 1$ falls $i > n$.

Für die Berechnung der Zustandswahrscheinlichkeiten von $M/M/c$-Systemen, also Stationen mit c Bedienern und unendlichem Warteraum gilt

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

und für die Wahrscheinlichkeit $p(0)$ gilt

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

mit $\rho = \lambda/(c \cdot \mu)$. Für eine Herleitung der Formeln, wie auch der Formeln des im Folgenden kurz vorgestellten $M/M/1/N$-System sei auf die Literatur wie z.B. [10] verwiesen.

Abbildung 3.7 zeigt den Verlauf der Population für verschiedene $M/M/c$-Systeme. Die Unterschiede sind relativ gering, mit einer steigenden Bedienerzahl erhöht sich die Population, da bei weniger als c Aufträgen an der Station die Bedienkapazität nicht voll ausgeschöpft werden kann.

Als letztes Modell wird kurz das $M/M/1/N$-System betrachtet. Ankünfte erfolgen mit Rate λ, solange weniger als N Aufträge im an der Station sind und Aufträge werden mit Rate μ bedient, solange mindestens ein Auftrag an der Station ist. Auch hier gilt natürlich wieder die rekursive Beziehung $p(i) = p(0) \cdot \rho^i$ und

$$p(0) = \left(1 + \sum_{i=1}^{N} \rho^i\right)^{-1} \;\Rightarrow\; p(0) = \begin{cases} \frac{1-\rho}{1-\rho^{N+1}} & \text{falls } \rho \neq 1 \\ \frac{1}{N+1} & \text{falls } \rho = 1 \end{cases}$$

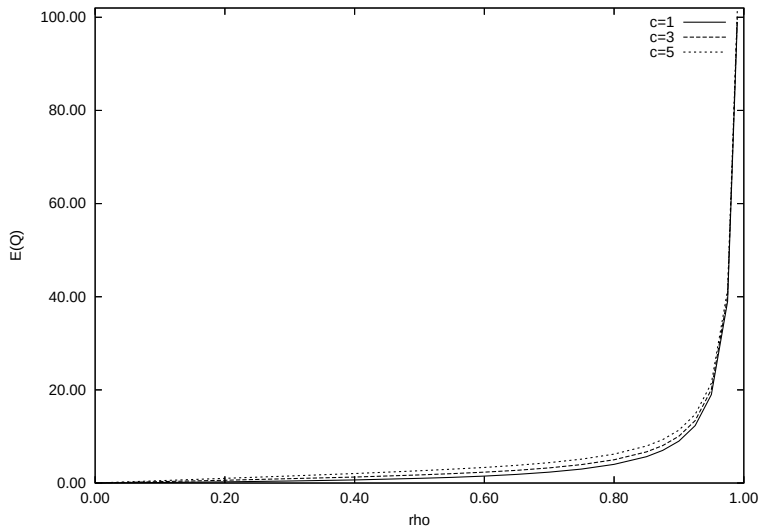

Abbildung 3.7: Verlauf der Population für verschiedene $M/M/c$-Systeme.

Da das System eine endliche Kapazität hat, kann der Wert von ρ auch größer als 1 sein. Der stationäre Zustand wird immer erreicht, da Aufträge bei vollem Warteraum einfach abgewiesen werden. Die Wahrscheinlichkeit einer Abweisung entspricht $p(N) = p(0) \cdot \rho^N$.

Es gibt zahlreiche weitere Modell vom Typ $M/M/\ldots$ für die Lösungen berechnet werden können. Das Vorgehen ist dabei immer identisch und besteht aus den folgenden 3 Schritten.

1. Aufstellen der rekursiven Beziehungen zwischen den Zustandswahrscheinlichkeiten (für Systeme mit Einzelankünften und Einzelbedienung als Geburts-/Todesprozess).
2. Berechnung der Zustandswahrscheinlichkeiten.
3. Ermittlung der Leistungsgrößen.

Je nach Art des Systems können mehr oder weniger explizite Berechnungen hergeleitet werden. Im Prinzip lässt sich das Vorgehen auch auf komplexere Modelle übertragen, die dabei entstehenden allgemeinen Gleichungssysteme müssen aber in der Regel numerisch gelöst werden.

Stationen mit nicht exponentiell-verteilten Zeiten

Es gibt eine Vielzahl von Arbeiten, die sich mit Stationen vom Typ $G/M/1$, $M/G/1$ und $G/G/1$ beschäftigen. Die meisten Resultate für diese allgemeineren Stationstypen sind deutlich komplexer als die hier vorgestellten Ergebnisse und es gibt wenige explizite Resultate. Dies liegt darin begründet, dass die speziellen Eigenschaften der Exponentialverteilung, nämlich Additivität und Gedächtnislosigkeit verloren gehen. Eines der wenigen expliziten Resultate ist die bekannte Pollaczek-Khinchine Formel für die Population in $M/G/1$-Systemen mit $FCFS$ Bedienung. Es gilt dort

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

mit $VK^2(B) = \sigma^2(B)/(E(B))^2$ dem quadrierten Variationskoeffizienten der Bedienzeitverteilung. Da der Durchsatz der Ankunftsrate λ entspricht, kann mit Hilfe des Satzes von Little auch die Verweilzeit berechnet werden. Die Formel zeigt, dass mit steigender Variabilität der Bedienung auch die Population in der Station steigt. Außerdem ist die Population nur von den ersten zwei Momenten der Bedienzeit und nicht von der gesamten Verteilung abhängig. Dieser recht einfache Zusammenhang ist allerdings die Ausnahme, in allen anderen Fällen beeinflusst die gesamte Verteilung die Leistungsgrößen und die Betrachtung einzelner Momente reicht nicht aus.

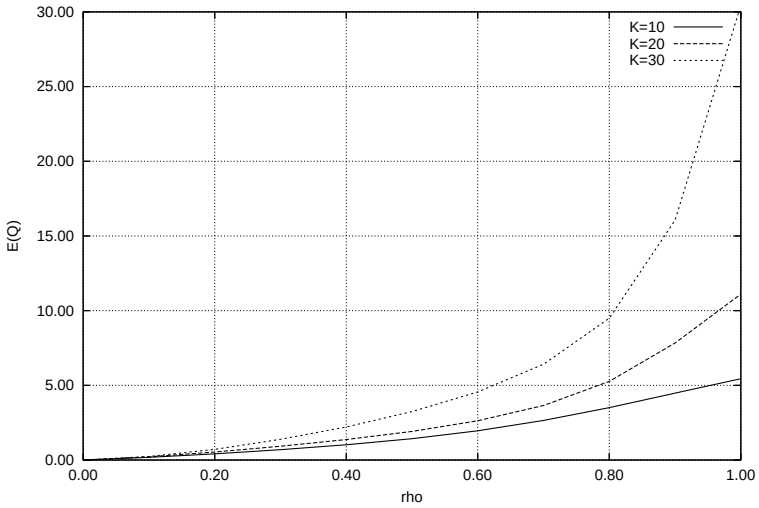

Abbildung 3.8: Verlauf der Population für $M/M/1/N$-Systeme.

3.2 Offene Warteschlangennetze

Die bisher betrachteten einzelnen Stationen erlauben die Modellierung eines Typs von Ressource, die von Aufträgen mit stochastisch identischen Bedienwünschen belegt wird. Reale Systeme bestehen aber aus einer Menge von Ressourcen, die von unterschiedlichen Auftragstypen belegt werden. Um solche Systeme adäquat zu modellieren muss die Modellklasse erweitert werden. Dies geschieht dadurch, dass Aufträge mehrere Stationen nacheinander oder alternativ durchlaufen. Zusätzlich werden, falls notwendig, noch unterschiedliche Auftragstypen, in der Regel als Auftragsklassen bezeichnet, eingeführt. Es gibt eine Vielzahl unterschiedlicher Klassen von Warteschlangennetzen. Grob unterschiedet man in geschlossene Modelle, die eine feste Anzahl Aufträge permanent beinhalten und offene Modelle, bei denen Aufträge das Modell betreten und auch wieder verlassen. Wir wollen hier kurz als einfachste Klasse die offenen Warteschlangennetze mit einer Auftragsklasse betrachten. Weitere Resultate und Anwendungsbeispiele findet man in [13]. Die hier betrachteten Warteschlangennetze sind charakterisiert durch

- J Stationen mit einem Bediener, unendlichem Warteraum und exponentieller Bedienzeitverteilung mit Rate μ_i an Station $i \in \{1, \ldots, J\}$.
- Aufträgen, die mit exponentiell verteilten Zwischenankunftszeiten und Rate λ_{0i} an Station i ankommen.
- Routingwahrscheinlichkeiten p_{ij}, die angeben, dass ein Auftrag, der Station i verlässt, mit Wahrscheinlichkeit p_{ij} Station j als nächstes besucht, falls $j \in \{1, \ldots, J\}$ und mit Wahrscheinlichkeit p_{i0} das Modell verlässt. Es muss gelten $\sum_{j=0}^{J} p_{ij} = 1$ für alle $i \in \{1, \ldots, J\}$.

Sei λ_i die Ankunftsrate an Station i, die Ankünfte von außen und von anderen Stationen umfasst. Damit stationäre Gleichgewicht herrscht muss $\lambda_i < \mu_i$ für alle Stationen im Netz gelten. Für die Ankunftsrate λ_i gilt

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^{J} p_{ji} \cdot \lambda_j \ .$$

Dies kann man in Vektor-Matrix-Schreibweise als

$$\Lambda = \Lambda P + \Lambda_0 \tag{3.3}$$

darstellen. Dabei ist P eine $J \times J$-Matrix mit p_{ij} an Position i, j $(1 \leq i, j \leq J)$, $\Lambda_0 = (\lambda_{01}, \ldots, \lambda_{0J})$ und $\Lambda = (\lambda_1, \ldots, \lambda_J)$. Vektor Λ_0 ist bekannt und Vektor Λ ist zu berechnen. Ein einfaches Umformen des Gleichungssystems liefert

$$\Lambda = \Lambda_0 \cdot (I - P)^{-1}$$

Die Lösung existiert also, wenn die inverse Matrix existiert. Die inverse Matrix existiert, wenn sich die Zeilen und Spalten von P sich nicht durch symmetrische Permutationen so umordnen lassen, dass eine Matrix der Form

$$\begin{pmatrix} A & 0 \\ C & B \end{pmatrix}$$

entsteht, bei der alle Zeilensummen von A gleich 1 sind.Für den fall, dass die inverse Matrix existiert sei $N = (I - P)^{-1}$. N ist eine nicht negative Matrix (siehe [8]).

Die Berechnung der Ankunftsraten erfordert damit die Lösung eines Gleichungssystems der Dimension J oder die Berechnung der inversen Matrix. Die Inversenbildung empfiehlt sich nur, wenn Matrix N zu späteren Resultatberechnung benutzt wird. Zur Analyse des Netzes müssen natürlich die Resultate für die Stationen und für die Aufträge, die das System durchlaufen, berechnet werden.

Der Zustand des Netzes ist gegeben durch einen Vektor $(n_1, \ldots, n_J)$, wobei n_i die aktuelle Population an Station i ist. Sei nun

$$\rho_i = \lambda_i / \mu_i \text{ und } \quad P[n_i] = (1 - \rho_i) \cdot (\rho_i)^{n_i}$$

Die Zustandswahrscheinlichkeiten entsprechen gerade den Zustandswahrscheinlichkeiten eines $M/M/1$-Systems (siehe Gl. 3.2). Das folgende fundamentale Resultat von Jackson von 1963 [10] erlaubt die Berechnung der Zustandswahrscheinlichkeit für einen Netzzustand aus den Stationszustandswahrscheinlichkeiten. Es gilt

$$P[n_1, \ldots, n_J] = \prod_{i=1}^{J} P[n_i] \ .$$

Auch wenn dieses Resultat auf den ersten Blick nicht überraschen mag, so stellt sich bei näheren Untersuchungen schnell heraus, dass seine Gültigkeit nicht unbedingt selbstverständlich ist. Die obige Darstellung des Gesamtzustandes, die man auch als Produktform bezeichnet, tritt nur bei einigen speziellen Warteschlangennetzen auf. Schon bei den hier betrachteten Jackson-Netzen sind die Ankunftsströme an den einzelnen Stationen nicht zwangsläufig Poisson-Prozesse. Immer dann, wenn Aufträge Stationen mehrfach durchlaufen könne, also so genannte Zyklen im Netz entstehen, bilden die Ankünfte sehr komplexe stochastische Prozesse, die nicht einfach beschreibbar und noch weniger analysierbar sind. Trotzdem verhalten sich die Stationen so, als wären alle Ankunftsprozesse Poisson-Prozesse und als wären alle Stationen unabhängig.

Auf Basis der Produktform lassen sich die Zustandswahrscheinlichkeiten des Netzes einfach berechnen:

1. Berechnung der Ankunftsraten λ_j durch Lösung eines linearen Gleichungssystems.
2. Analyse von J $M/M/1$-Systemen.

Aus der Analyse resultieren die Leistungsgrößen $E(U_i)$, $E(X_i)$, $E(Q_i)$ und $E(R_i)$ für jede einzelne Station im Netz. Um Resultate für das gesamte Netz zu bestimmen, muss man die einzelnen Größen entsprechend zusammenfassen. Der Durchsatz entspricht bei Stationarität gerade der Summe der Ankunftsraten

$$E(X_0) = \sum_{i=1}^{J} \lambda_{0i}$$

und die Population entspricht der Summe der Populationen in den Stationen

$$E(Q_0) = \sum_{i=1}^{J} E(Q_i)$$

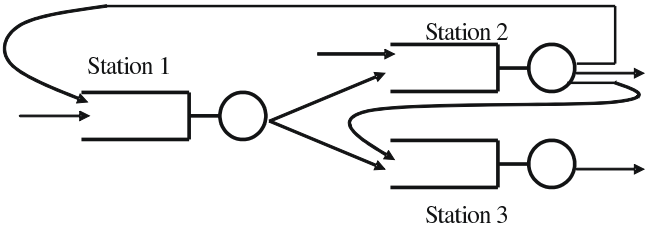

Abbildung 3.9: Einfaches Jackson-Netz.

Damit kann der Satz von Little auf das gesamte Netz angewendet werden und es gilt

$$E(R_0) = E(Q_0)/E(X_0)$$

Etwas komplexer ist die Berechnung des Erwartunsgwertes der Zeit, die ein zufällig gewählter Auftrag an Station j verbringt, bevor er das Netz verlässt. Sei

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

die Wahrscheinlichkeit, dass ein zufällig gewählter Auftrag das Netz in Station i betritt. Wie in [8] gezeigt wird, ist $N(i,j)$ die mittlere Anzahl von Besuchen in Station j vor Verlassen des Netzes für einen Auftrag der sich in Station i befindet. Damit ergibt sich die mittlere Verweilzeit an Station j als

$$\sum_{i=1}^{J} p_{0i} \cdot N(i,j) \cdot E(R_j)$$

Das in Abbildung 3.9 gezeigte einfache Beispiel soll nun analysiert werden. Dazu seien die folgenden Parameter gegeben

$$\mu_1 = \mu_2 = \mu_3 = 1.0$$
$$\lambda_{01} = 0.4,\ \lambda_{02} = 0.2$$
$$p_{12} = p_{13} = p_{23} = 0.5,\ p_{21} = 0.4,\ p_{20} = 0.1,\ p_{30} = 1.0$$

Alle weiteren Parameterwerte sind 0. Die Ankunftsraten werden aus dem folgenden Gleichungssystem ermittelt.

$$\left.\begin{array}{l} \lambda_1 = 0.4 + 0.4 \cdot \lambda_2 \\ \lambda_2 = 0.2 + 0.5 \cdot \lambda_1 \\ \lambda_3 = 0.5 \cdot \lambda_1 + 0.5 \cdot \lambda_2 \end{array}\right\} \Rightarrow\ \lambda_1 = 0.6,\ \lambda_2 = 0.5 \ \text{ und } \ \lambda_3 = 0.55$$

Damit sind $M/M/1$-Systeme mit $\rho = 0.6, 0.5, 0.55$ zu analysieren. Es gilt $E(R_1) = 2.5$, $E(R_2) = 2.0$, $E(R_3) = 2.222$ und $E(Q_i) = E(R_i) - 1$ für alle i. Die Wahrscheinlichkeit dafür, dass das gesamte Netz keine Aufträge enthält lautet

$$P[0,0,0] = P[n_1 = 0] \cdot P[n_2 = 0] \cdot P[n_3 = 0] = 0.4 \cdot 0.5 \cdot 0.45 = 0.09$$

Wenn $\lambda_{01} = 2 \cdot \lambda_{02}$ gesetzt wird und $\lambda_0 = \lambda_{01} + \lambda_{02}$ definiert wird, so kann man die Verweilzeit für steigende Ankunftsrate λ_0 bestimmen. Wie Abbildung 3.10 zeigt, wächst dabei die Verweilzeit am Flaschenhals (d.h. der am höchsten ausgelasteten Station) am schnellsten und bestimmt für hohe Last die Verweilzeit im Netz maßgeblich.

Die hier vorgestellten Jackson-Netze markieren nur einen Anfang in der Analyse von Warteschlangennetzen. Es gibt inzwischen eine Vielzahl von Resultaten, deren Vorstellung über die Ziele dieser Vorlesung hinausgehen würden. Es sollen nur kurz die folgenden wesentlichen Aspekte erwähnt werden.

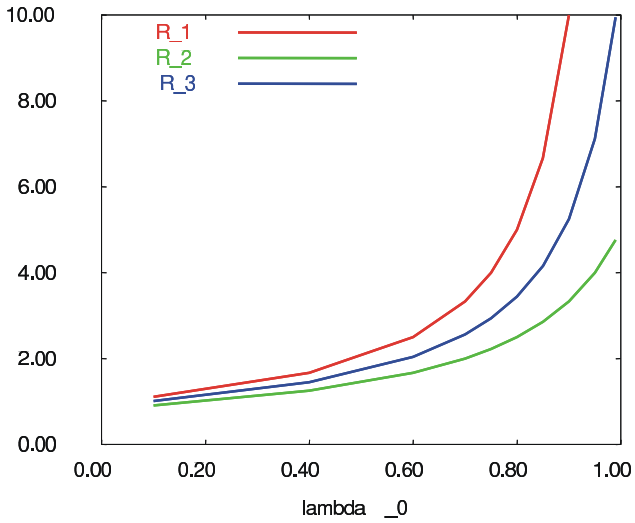

Abbildung 3.10: Verlauf der Verweilzeit im Beispielmodell.

- Der vorgestellte Analyseansatz kann einfach erweitert werden, wenn Stationen lastabhängige Bedienraten haben. In diesem Fall, der auch Mehrbedienerstationen umfasst, müssen nur die zugehörigen Analysen auf Stationsebene integriert werden.
- Die Integration von Stationen endlicher Kapazität oder mit nicht-exponentiellen Bedienzeitverteilungen ist nicht möglich, ohne die Produktform Eigenschaft zu verlieren.
- Es lassen sich mehrere Auftragsklassen mit unterschiedlichem Verhalten integrieren, dann wird allerdings die Bedienstrategie an den Stationen bedeutsam.
- Weiterhin lassen sich die Ansätze auf bestimmte geschlossene Netze übertragen, bei denen einen endliche Auftragszahl permanent im Netz zirkuliert.