

Lehrstuhl für Informatik IV
Rheinisch-Westfälische Technische Hochschule Aachen
Prof. Dr. Otto Spaniol

ANALYTISCHE LEISTUNGSBEWERTUNG
EINER PARALLELEN CONTROLLER-ARCHITEKTUR
FÜR HOCHGESCHWINDIGKEITSPROTOKOLLE

Diplomarbeit

vorgelegt von
cand. inform. Andreas Grün
Matrikelnummer 155037

Aachen, den 4. März 1994

Betreuung:
Prof. Dr. Otto Spaniol
Dipl. Inform. Christian Engel

ERKLÄRUNG

Hiermit versichere ich, daß ich diese Diplomarbeit selbständig verfaßt und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Aachen, den 4. März 1994

Andreas Grün

Inhaltsverzeichnis

1. Einleitung	1
1.1 Motivation	1
1.2 Das Ziel der Arbeit	2
2. Leistungsbewertung	3
2.1 Anwendungsgebiete für die Leistungsbewertung	3
2.2 Leistungsgrößen	4
2.3 Die Methoden zur Leistungsbewertung	4
2.3.1 Meßmethoden	4
2.3.2 Modellbildungstechniken	5
2.4 Stochastische Methoden bei der analytischen Leistungsbewertung	7
2.5 Literatur zur Leistungsbewertung	9
3. Die PAPER-Architektur	10
3.1 Das funktionale Konzept der PAPER-Architektur	11
3.2 Die Elemente der PAPER-Architektur	11
3.2.1 Die Kontrolleinheit	12
3.2.2 Die Ausführungseinheit	14
3.2.3 Die Verbindung von Kontroll- und Ausführungseinheit	14
4. Warteschlangentheorie	16
4.1 Die Beschreibung von Warteschlangensystemen	16
4.1.1 Kendall'sche Notation	18
4.2 Wahrscheinlichkeitstheoretische Grundlagen	19
4.2.1 Zufallsvariablen	19
4.2.2 Wahrscheinlichkeitsverteilungen	23
4.2.3 Stochastische Prozesse	25
4.3 Elementare Wartesysteme	33
4.3.1 Leistungsgrößen eines Wartesystems	34
4.3.2 Das Gesetz von Little	35

4.3.3	M/M/1-Wartesysteme	37
4.3.4	M/M/c-Wartesysteme	39
4.3.5	M/M/c-Wartesysteme mit statischer Prioritäts-Disziplin	43
4.4	Warteschlangennetze	50
4.4.1	Formale Beschreibung von Warteschlangennetzen	51
4.4.2	Produktform-Warteschlangennetze	52
5.	Das Warteschlangenmodell der PAPER- Architektur	57
5.1	Die Kunden des Warteschlangenmodells	57
5.2	Die Konstruktion eines Warteschlangennetzes für die PAPER-Architektur	58
5.2.1	Die Modellierung der Kontrolleinheit	58
5.2.2	Die Modellierung der Ausführungseinheit	60
5.2.3	Die Modellierung der Verbindung von Kontroll- und Ausführungseinheit	62
5.3	Die Berechnung der Ankunftsraten und Verzweigungswahrscheinlichkeiten	63
5.3.1	Die Berechnung von $\vec{\lambda}$ für ein bekanntes λ_{02}	65
5.3.2	Die Berechnung von $\vec{\lambda}$ für $\lambda_{02} = \gamma \lambda_1$	68
5.4	Die Berechnung der Verzweigungswahrscheinlichkeiten p_{54} und p_{56}	71
6.	Die Analyse des Warteschlangenmodells	73
6.1	Die Analyse der Ankunftsraten	73
6.2	Die Analyse der Verzweigungswahrscheinlichkeiten	75
6.2.1	Die Verzweigungswahrscheinlichkeiten des Markenspeichers	76
6.2.2	Die Verzweigungswahrscheinlichkeiten des Deaktivierers	77
6.2.3	Die Verzweigungswahrscheinlichkeiten der Sperreinheit	78
6.2.4	Die Verzweigungswahrscheinlichkeiten der Schaltmaschinen	78
6.3	Die Ankunftsraten der Prioritätsklassen	80
7.	Die Bewertung der PAPER-Architektur durch das Warteschlangenmodell	83
7.1	Die benötigten Basisgrößen für die Bewertung	83
7.2	Die Dimensionierung der Verbindungswarteschlangen	89
7.3	Die Anzahl der Bedieneinheiten für die Knoten des Warteschlangennetzes	92
7.3.1	Die Anzahl der Schaltmaschinen	92
7.3.2	Die Bedieneinheiten weiterer Knoten	95
7.4	Das Verhältnis von Aktivierungs- und Schaltzeit	98

7.4.1	Die Schaltzeit	98
7.4.2	Die Aktivierungszeit	100
7.4.3	Vergleich und Bewertung von Aktivierungs- und Schaltzeit	101
7.5	Die Umlaufzeit einer Transition	102
7.6	Zusammenfassung der Analyse	103
8.	Zusammenfassung und Ausblick	105
	Literaturverzeichnis	107
A.	Die Leistungsgrößen von Wartesystemen	I
B.	Die Ankunftsraten und Verzweigungswahrscheinlichkeiten	III

1. Einleitung

1.1 Motivation

Die Weiterentwicklung der Rechnerarchitekturen hat im Verlauf der letzten Jahre zu immer leistungsfähigeren Rechensystemen geführt, die in fast allen Bereichen des täglichen Lebens zum Einsatz kommen. Die vielfältigen Anwendungsgebiete von Rechensystemen machte die Vernetzung zwischen den Systemen notwendig.

Dabei stand zunächst der Wunsch nach einer gemeinsamen Nutzung von teuren Geräten im Vordergrund. Später kam dann die Nutzung von zentralen Datenbeständen und die Kommunikation untereinander hinzu. Das führte zur Entwicklung von leistungsfähigen Übertragungsmedien, die versuchten, den steigenden Ansprüchen der Benutzer gerecht zu werden. So sind bei speziell entwickelten Hochgeschwindigkeitsnetzen (z.B. FDDI¹) Übertragungsraten von ca. 100 Mbps² möglich.

Um solche hohen Übertragungsleistungen nutzen zu können und dem immer höher werdenden Datenaufkommen gerecht zu werden, sind entsprechend leistungsfähige Rechner und Kommunikations-Controller notwendig. Letzteren kommt bei der Datenübertragung eine besondere Rolle zu. Zum einen müssen sie die Nutzdaten senden und empfangen. Zum anderen fällt ihnen die Aufgabe der Protokollverwaltung zu, was eine zusätzliche Belastung bedeutet. Dabei hat es sich gezeigt, daß die Kommunikationsprotokolle nicht in der Lage sind, die hohen Anforderungen von z.B. Multimedia-Anwendungen zu erfüllen. Es zählt also zu den vorrangigen Zielen, sowohl die einzusetzende Controller-Hardware, wie auch die darauf zu implementierenden Kommunikationsprotokolle zu verbessern. Ein Lösungsansatz ist die *Parallelisierung* der Kommunikationsprotokolle mit Hilfe von Petri-Netzen.

Das Forschungsprojekt PIKOM³ am Lehrstuhl für Informatik IV der RWTH Aachen befaßt sich mit der Spezifikation, Implementierung und Leistungsbewertung von Kommunikationsprotokollen unter dem Aspekt der Parallelverarbeitung. Dabei bilden Petri-Netze die Basis für die formale Beschreibung und die Implementierung der Kommunikationsprotokolle. Zur Beschreibung paralleler Vorgänge wurden bei der GMD in Darmstadt die Produktnetze als spezielle Form von höheren Petri-Netzen entwickelt [BOP89]. Auf der Basis von durch Produktnetzen beschriebenen Protokollen wurde in mehreren Arbeiten ([Rup87], [Eng90]) eine parallele Controller-Architektur entwickelt, aus der die PAPER⁴-Architektur als augenblickliches Entwicklungsergebnis hervorgeht [Eng94].

¹ **Fiber Distributed Data Interface**

² **mega bits per second**

³ **Parallelität in Kommunikationsprotokollen**

⁴ **Petri Net Based Parallel Architecture for Protocol Engineering and Realization**

Im Rahmen der Entwicklung der PAPER-Architektur wurde die Programmiersprache PENCIL/C entwickelt [Eng92]. Sie stellt eine Erweiterung von ANSI-C dar und bietet mittels eines speziellen Compilers [Gro93] die Möglichkeit, in PENCIL/C beschriebene Kommunikationsprotokolle auf der PAPER-Architektur zu implementieren.

Bei der PAPER-Architektur handelt es sich um eine *abstrakte Architektur* zum Entwurf eines Mehrprozessor-Kommunikations-Controllers, die eine Realisierung auf den verschiedensten Hardware-Plattformen (z.B. Transputer-Cluster) möglich macht. Gerade in der Entwurfsphase einer Hardware-Architektur ist zu prüfen, ob sie den gestellten Anforderungen genügt. Dabei ist es notwendig Einzelkomponenten so auszuwählen, daß sie in ihrer Leistungsfähigkeit aufeinander abgestimmt werden können. Auch müssen Realisierungsalternativen bewertet werden, um Entscheidungshilfen zu gewinnen. Die Leistungsbewertung hilft somit beim Systementwurf kostspielige Fehlentwicklungen zu vermeiden.

1.2 Das Ziel der Arbeit

Die vorliegende Diplomarbeit ist Teil des Forschungsprojekts PIKOM. Ihr Ziel ist die analytische Leistungsbewertung der PAPER-Architektur. Zusammen mit einem im Rahmen von PIKOM entwickelten PAPER-Emulator sind so Aussagen über das Leistungsverhalten der PAPER-Architektur möglich, die eine spätere Hardware-Realisierung unterstützen.

Das Kapitel 2 gibt ein Überblick über die Anwendungsgebiete und Methoden der Leistungsbewertung. Der Schwerpunkt liegt dabei auf der Vorgehensweise bei den analytischen Methoden. Im Kapitel 3 wird das funktionale Konzept der PAPER-Architektur beschrieben und deren Komponenten vorgestellt.

Die Leistungsbewertung der PAPER-Architektur erfolgt mit Hilfe der Warteschlangentheorie. Die warteschlangentheoretischen Grundlagen, die für diese Arbeit von Bedeutung sind, werden im Kapitel 4 eingeführt.

Im Rahmen der analytischen Leistungsbewertung spielt die Modellbildung eine zentrale Rolle. So ist das Modell ein Abbild der Realität. Die Abstraktion der PAPER-Architektur in die Modellwelt der Warteschlangen erfolgt im Kapitel 5. Es umfaßt die Konstruktion eines Warteschlangennetzes und die Herleitung der Berechnungsvorschriften für die spätere Analyse.

Ausgehend von dem Warteschlangenmodell wird dessen Verhalten im Kapitel 6 analysiert. Dem schließt sich die Analyse der Leistungsfähigkeit der PAPER-Architektur im Kapitel 7 an. Dabei werden einzelne Komponenten, sowie deren Zusammenwirken betrachtet. Als Basis für die Analyse dienen die auf einem PAPER-Emulator ermittelten Werte für eine Teilimplementierung des Kommunikationsprotokolls XTP. Die damit durchgeführten Berechnungen lassen Aussagen über das Leistungsverhalten der PAPER-Architektur zu.

Eine Zusammenfassung der Arbeit, sowie die wichtigsten Ergebnisse und Ansätze für weitere Arbeiten sind im Kapitel 8 zu finden. Im Anhang sind die in dieser Arbeit verwendeten Berechnungsvorschriften für Warteschlangensysteme und das Warteschlangennetz der PAPER-Architektur tabellarisch aufgelistet.

2. Leistungsbewertung

Die Leistungsbewertung von Rechen- und Kommunikationssystemen (kurz: *Systeme*) läßt sich definieren als die *quantitative Bestimmung der Leistungsfähigkeit* eines solchen Systems [Bol89]. Das dynamische Ablaufgeschehen innerhalb und zwischen den einzelnen Komponenten eines Systems soll untersucht und optimiert werden. Die Ziele der Leistungsbewertung sind

- die Messung und formale Beschreibung realer Systemabläufe,
- die Definition, Bestimmung und Analyse charakteristischer Leistungsgrößen (Bewertungsparameter),
- die Bereitstellung von Entscheidungshilfen.

2.1 Anwendungsgebiete für die Leistungsbewertung

Mit den oben genannten Zielsetzungen gibt es drei Anwendungsgebiete für die Leistungsbewertung [Lan92]:

1. Systementwurf

Beim Entwurf eines neuen Systems muß während der Entwurfsphase ständig überprüft werden, ob das System den gestellten Anforderungen genügt. Dabei ist es oftmals erforderlich, Einzelkomponenten so auszuwählen, daß sie in ihrer Leistungsfähigkeit aufeinander abgestimmt werden. Es müssen auch Realisierungsalternativen bewertet werden um Entscheidungskriterien gewinnen zu können. Die Leistungsbewertung als Leistungsvorhersage hilft somit schon in der Entwurfsphase kostspielige Fehlentwicklungen zu vermeiden.

2. Kontrolle und Tuning realer Systeme

Die Überwachung des Systemverhaltens und das Aufdecken von Engpässen ist die vorrangige Aufgabe der Leistungsbewertung bei realen Systemen. Die gezielte Beseitigung von Engpässen wird als *Tuning* bezeichnet. Mit den Methoden der Leistungsbewertung können Systemkomponenten, die das Leistungsverhalten des gesamten Systems negativ beeinflussen, ermittelt und Maßnahmen zu deren Beseitigung in die Wege geleitet werden.

3. Systemvergleich

Bei Neuanschaffungen werden Bewertungsgrundlagen benötigt, um aus mehreren Alternativen das geeignete System auswählen zu können. Das wesentliche Auswahlkriterium ist die Frage, ob ein System die erwartete Auftragslast bewältigen kann.

2.2 Leistungsgrößen

Um Aussagen über die Leistungsfähigkeit eines Systems machen zu können, müssen bestimmte Parameter definiert und analysiert werden. Diese Parameter werden *Leistungsgrößen* oder *Bewertungsparameter* genannt. Leistungsgrößen können auf vielfältige Weise definiert werden. Welche Leistungsgrößen in einem konkreten Anwendungsfall benötigt werden, hängt von der jeweiligen Zielsetzung der Leistungsbewertung und dem Anwendungsgebiet ab. In [Lan92] ist eine detaillierte Übersicht über die Leistungsgrößen von einzelnen Systemkomponenten und von gesamten Systemen im Hinblick auf ihre Anwendungsgebiete enthalten.

2.3 Die Methoden zur Leistungsbewertung

Leistungsbewertung wird mit unterschiedlichen Zielen durchgeführt. Daher kommen auch verschiedene Bewertungsmethoden zur Anwendung.

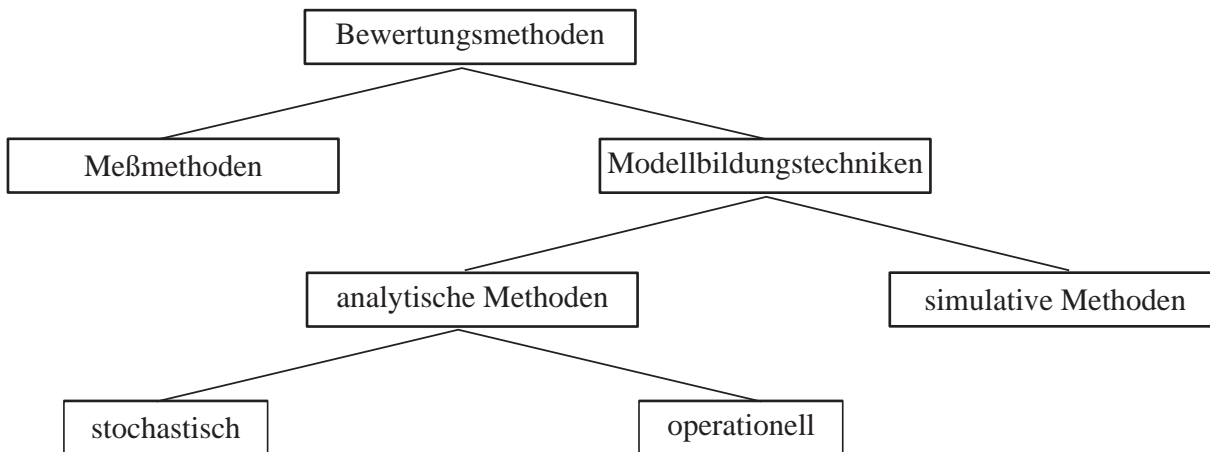


Abbildung 2.1: Die verschiedenen Methoden zur Leistungsbewertung im Überblick

2.3.1 Meßmethoden

Die Leistungsfähigkeit *real existierender* Systeme kann durch Meßmethoden ermittelt werden.

Zur Leistungsmessung eines Systems unter einer bestimmten Auftragslast werden *Benchmark-Tests* durchgeführt. Die Auftragslast ist die Menge aller Rechenvorgänge und Daten, die während eines Beobachtungszeitraumes im System verarbeitet werden. Sie hat einen großen Einfluß auf die zu bestimmenden Leistungsgrößen, wie z.B. die Auslastung oder die Antwortzeit einer Systemkomponenten oder eines gesamten Systems. In den meisten Fällen ist die aktuelle Auftragslast unvorhersehbaren Schwankungen unterworfen und folgt statistischen Gesetzmäßigkeiten.

Die Erfassung der Aktivitäten innerhalb eines Systems erfolgt durch *Leistungsmonitore*. Die Aktivitäten des Systems werden dabei erfaßt und in periodischen Abständen analysiert. Monitore sind unabhängige Meßgeräte und lassen sich in Hardware- und Software-Monitore unterteilen

[Lan92]. *Hardware-Monitore* sind, wie der Name vermuten läßt, Hardware-Bausteine, die innerhalb eines Systems elektrische Impulse messen. Sie erfassen und werten spezielle Systemdaten aus, die von der Hardware des zu analysierenden Systems abgenommen werden. Im Gegensatz dazu stehen die *Software-Monitore*. Diese bestehen aus einem oder mehreren speicherresidenten Programmen. Sie sind entweder Teil des Betriebssystems oder Anwenderprogramme mit besonderen Privilegien, die nach Ablauf einer bestimmten Zeit oder beim Eintreffen eines bestimmten Ereignisses Daten erfassen, aufbereiten und ausgeben.

Vorausgesetzt, daß die Messungen korrekt und über einen ausreichend langen Zeitraum durchgeführt werden, lassen sich durch diese Methode sehr zuverlässige Leistungsgrößen des Systems ermitteln.

Befindet sich das System noch in der Entwurfsphase oder sind Untersuchungen an einem realen System zu aufwendig und kostenintensiv, dann erfolgt die Leistungsbewertung durch Modellbildungstechniken.

2.3.2 Modellbildungstechniken

Die Leistungsbewertung eines realen Systems ist wegen dessen Komplexität oftmals zu schwierig oder das zu bewertende System befindet sich noch in der Entwicklungsphase. Um aber dennoch Aussagen bezüglich der Leistungsfähigkeit machen zu können, bedient man sich eines *Modells*. In dem Modell werden das dynamische Ablaufgeschehen und die für die Analyse relevanten Systemmerkmale nachgebildet. Die Systeme werden so weit abstrahiert, daß die interessanten Größen noch hinreichend gut erfaßt werden und irrelevante Details weitgehend unterdrückt werden. Die Betrachtung eines Modells bietet die folgenden Vorteile:

1. Die Entwicklung eines neuen Systems wird unterstützt, da in der Anfangsphase der Entwicklung noch keine Messungen möglich sind.
2. Alternative Systemkonfigurationen können miteinander verglichen werden. Eine Änderung des Modells ist meistens einfacher als die Änderung eines realen Systems.
3. Das Verhalten eines Systems unter geänderten äußeren Bedingungen kann untersucht werden.

Bei der Erstellung eines Modells muß darauf geachtet werden, daß

- alle wichtigen Systemeigenschaften erfaßt werden,
- unwichtige Systemeigenschaften nicht im Modell enthalten sind,
- das Modell eine gültige Darstellung der Realität, d.h. eines existierenden Systems oder eines zu entwickelnden Systems, ist.

Die Modellbildungstechniken können *analytisch* oder *simulativ* sein.

2.3.2.1 Simulative Modellbildungstechniken

Bei der Simulation wird das Systemverhalten experimentell untersucht und mit speziellen Computer-Programmen nachvollzogen. Oftmals wird dazu eine eigene Simulationssprache verwendet. Da das Verhalten eines Simulationsmodells in Bezug auf die relevanten Systemparameter dem Verhalten des zu modellierenden Systems entspricht, können daraus alle zur Leistungsbeurteilung interessanten Größen abgeleitet und ermittelt werden.

Bei der simulativen Modellbildungstechnik lassen sich drei Verfahren unterscheiden:

1. Monte-Carlo-Simulation

Bei der Monte-Carlo-Simulation wird die Leistungsanalyse nach Gesetzmäßigkeiten der Wahrscheinlichkeitstheorie durchgeführt. Um eine bestimmte Leistungsgröße zu analysieren, wird sie in ein wahrscheinlichkeitstheoretisches Modell eingebunden, mit dessen Hilfe sie dann abgeschätzt werden kann.

2. Ereignisorientierte Simulation

Die ereignisorientierte Simulation verwaltet eine Liste von definierten Ereignissen. Durch ein Ereignis E_{t_i} , welches zum Zeitpunkt t_i eintritt, wird in der Simulation nach definierten Regeln eine Zustandsänderung innerhalb des Modells ausgelöst. Durch die Zustandsübergänge treten weitere Ergebnisse ein, die das dynamische Ablaufgeschehen des modellierten Systems wiedergeben.

3. Vorgangorientierte Simulation

Im Gegensatz zur ereignisorientierten Simulation, bei der Systemzustände definiert und durch Ereignisse beschrieben werden müssen, wird bei der vorgangorientierten Simulation das dynamische Ablaufgeschehen aus der Sicht einer zu bearbeitenden Last beschrieben. Es werden Lastzustände definiert, die den Weg der Last durch das System wiedergeben. Zu jedem Zeitpunkt befindet sich die Last in einem bestimmten Zustand.

Simulative Modellbildungstechniken liefern auch Ergebnisse für Modelle, die wegen ihrer Komplexität nicht durch analytische Methoden erfaßbar sind. Hierbei ist allerdings der Programmieraufwand und die zu erwartende Laufzeit des Simulationsprogramms nicht zu unterschätzen. In [Jai91] findet der interessierte Leser eine detaillierte Einführung in die Leistungsbewertung mittels simulativer Modellbildungstechniken.

2.3.2.2 Analytische Modellbildungstechniken

Bei der analytischen Leistungsbewertung werden auf mathematischem Weg Beziehungen zwischen fundamentalen Systemgrößen und relevanten Leistungsgrößen hergeleitet. Dabei entsteht ein Spannungsfeld zwischen der Genauigkeit des Modells und dessen mathematischer Handhabbarkeit. Analytische Modelle lassen sich in *stochastische* und *operationelle* Modelle unterteilen.

Stochastische Modelle: Bei den stochastischen Modellen sind die Systemparameter (z.B. Bearbeitungszeiten oder Ankunftszeiten) als Zufallsvariablen und das Systemverhalten durch

stochastische Prozesse modelliert. Die Ergebnisse der Analyse sind dementsprechend auch statistisch verteilte Leistungsgrößen. Die Systembeschreibung erfolgt in einer Modellwelt, die sich wahrscheinlichkeitstheoretischer Beschreibungstechniken bedient.

Operationelle Modelle: Bei den operationellen Modellen werden keine wahrscheinlichkeitstheoretischen Konzepte für die Modellierung der Systemparameter benutzt. Das Systemverhalten wird während eines festen Zeitintervalls beobachtet, wobei einfach zu erfassende Basisgrößen gemessen werden. Aus diesen Basisgrößen werden anschließend durch einfache mathematische Verfahren andere Leistungsgrößen abgeleitet, deren Messung zu aufwendig wäre.

Die Beschränkung auf ein festes Beobachtungsintervall führt zu einer wesentlich einfacheren Bestimmung der Leistungsgrößen als bei der stochastischen Modellbildung. Deren Vorteil sind jedoch einfache und schnelle Algorithmen und leicht zu interpretierende Beziehungen zwischen den Modellparametern und den relevanten Leistungsgrößen. Es ist aber oft schwierig, für komplexe Systeme exakte analytische Modelle zu bilden.

Bei der in dieser Diplomarbeit zu bewertenden PAPER-Architektur handelt es sich um das Konzept einer parallelen Kommunikations-Controller-Architektur, für das noch keine Hardware-Realisierung existiert. Aus diesem Grund ist eine Leistungsbewertung nur durch *Modellbildungstechniken* möglich. Eine Untersuchung des Leistungsverhaltens auf der Basis der simulativen Modellbildungstechnik wurde im Rahmen des Forschungsprojekts PIKOM durch die Entwicklung eines Emulators für die PAPER-Architektur durchgeführt. In der vorliegenden Arbeit erfolgt die Leistungsbewertung der PAPER-Architektur auf der Abstraktionsebene der analytischen Modellbildungstechnik. Da die Abläufe innerhalb der Architektur zum jetzigen Zeitpunkt durch ein hohes Maß an *Unbestimmtheit* charakterisiert sind, wird dabei die stochastische Methode eingesetzt. Diese wird im folgenden Abschnitt näher erläutert.

2.4 Stochastische Methoden bei der analytischen Leistungsbewertung

Um bei der analytischen Leistungsbewertung verlässliche Aussagen mit einem vertretbaren Aufwand machen zu können, kommt der *Modellerstellung* eine zentrale Bedeutung zu. Dabei soll ein Modell die Eigenschaften eines realen Systems so genau wie möglich erfassen, andererseits muß das Modell noch mathematisch handhabbar sein.

Es gibt verschiedene Alternativen zur Abbildung des Systemverhaltens in eine Modellwelt. Bei der stochastischen Leistungsbewertung werden häufig *Warteschlangenmodelle* oder *stochastische Petri-Netze* verwendet. Beide Modellierungsformen sind von einer Untermenge der stochastischen Prozesse abgeleitet: den *Markov-Prozessen* (⇨ Abschnitt 4.2.3). Das dynamische Ab-

laufgeschehen wird durch eine Menge von Zuständen und den zugehörigen Zustandsübergangswahrscheinlichkeiten dargestellt.

Der Vorteil der Modellierung mittels Warteschlangenmodellen oder stochastischen Petri-Netzen liegt in der Beschreibungsform der Systemkomponenten und deren Verhalten: Das Modell basiert in einem hohen Maße auf einer graphischen und nicht so sehr auf einer mathematischen Beschreibung. Bei der Analyse ist eine genaue Kenntnis der zugrunde liegenden mathematischen Theorie nicht unbedingt notwendig. Warteschlangenmodelle können unter bestimmten Voraussetzungen *direkt* berechnet werden, ohne das zugrunde liegende Markov-Modell zu lösen. Bei der Verwendung von stochastischen Petri-Netzen ist das bis jetzt nicht so einfach möglich [Pag86], da es sich bei diesen um ein sehr junges Hilfsmittel zur Modellierung handelt, das noch nicht so weit erforscht worden ist wie die Warteschlangentheorie.

Die Modellerstellung läßt sich in drei Schritten durchführen:

1. Konfigurationsbeschreibung
2. Lastbeschreibung
3. Modellbeschreibung

Bei der Modellerstellung kommt es zunächst darauf an, die wichtigsten, d.h. für die Leistungsbeurteilung interessanten Systemkomponenten quantitativ zu beschreiben und deren Interaktionen zu erfassen (*Konfigurationsbeschreibung*). Um das dynamische Ablaufgeschehen innerhalb des Systems bestimmen zu können, müssen die Gesetzmäßigkeiten bei der Abwicklung der Auftragslast identifiziert und beschrieben werden. Die Parameter für die *Lastbeschreibung* werden entweder gemessen, geschätzt oder aus anderen Größen hergeleitet. Daran zeigt sich, daß die Quantifizierung der Lastbeschreibung oftmals von Ungenauigkeiten geprägt ist. Auf der Grundlage der Konfigurations- und Lastbeschreibung kann die *modellhafte Erfassung des Ablaufgeschehens* erfolgen. Dieser Abstraktionsschritt erfolgt in eine Modellwelt, die die Basis für die *Modellbeschreibung* bildet. Wegen ihrer gut erforschten Handhabbarkeit ist die Modellwelt in vielen Fällen die Warteschlangentheorie. Diese wahrscheinlichkeitstheoretische Modellierungstechnik sollte verwendet werden [Her87], wenn

- komplexe deterministische Vorgänge mittels stochastischer Prozesse angenähert werden sollen,
- die zu beschreibenden Abläufe durch einen hohen Grad an Unbestimmtheit charakterisiert sind,
- nicht genügend Informationen über das Ablaufverhalten innerhalb des zu modellierenden Systems vorliegen.

Gerade in der Entwurfs- und Entwicklungsphase eines Systems treffen die drei genannten Gründe für eine wahrscheinlichkeitstheoretische Modellierungstechnik mittels Warteschlangen zu. Bei den Warteschlangenmodellen (⇨ Kapitel 4) wird die Lastbeschreibung bzw. das Ablaufgeschehen mittels stochastischer Prozesse beschrieben. Die Lastbeschreibung gliedert sich dabei in Angaben

über:

- den Ankunftsprozeß (die für jede typische Last durch Zufallsvariablen charakterisierten Ankunftszeitpunkte),
- den Bedienprozeß (die für jede typische Last und modellierte Systemkomponente durch Zufallsvariablen charakterisierte Bedienzeitanforderung).

Das Bediensystem, also die einzelnen Systemkomponenten, werden durch Angaben über die

- Art und Anzahl der Bedieneinheiten,
- Bedienzeiten für eine typische Last,
- Systemstruktur (Verbindungswege innerhalb des Modells),
- Bearbeitungsstrategie

charakterisiert.

Zusammenfassend läßt sich sagen, daß sich das Vorgehen für die analytische Leistungsbewertung im wesentlichen aus vier Vorgängen zusammensetzt:

1. *Modellbildung*
Abstraktion eines Systems in eine Modellwelt
2. *Analyse*
Systematische Untersuchung des Modells nach vorher definierten Zielen
3. *Bewertung*
Die durch die Analyse erhaltenen Ergebnisse werden in Bezug auf ihre Aussagefähigkeit bewertet und die Analyse u.U. mit neuen Startparametern wiederholt
4. *Übertragung* auf die Realität
Die erhaltenen Ergebnisse werden auf das reale System übertragen oder bei der weiteren Entwicklung berücksichtigt.

2.5 Literatur zur Leistungsbewertung

Einen umfassenden Überblick zum Thema Leistungsbewertung ist in dem Buch „*The Art of Computer Systems Performance Analysis*“ von R. Jain [Jai91] zu finden. Auch in [Lan92] und [Kob78] werden alle Bereiche der Leistungsbewertung umfassend dargestellt. [Mar90] beschreibt die stochastischen Methoden mittels Warteschlangen und stochastischer Petri-Netze. In [Pet81] und [Pag86] wird die analytische Leistungsbewertung durch stochastische Petri-Netze erläutert. Die operationelle Leistungsbewertung mittels Warteschlangen steht in [Laz84] im Vordergrund. In den Büchern [Kle75], [Kle76], [Bol89], [Kin90] und [Gro74] wird eingehend auf die Leistungsbewertung mittels Warteschlangenmodellen und deren Lösungen eingegangen.

3. Die PAPER-Architektur

Die PAPER¹-Architektur ist eine *abstrakte*, auf Petri-Netzen basierende Architektur zum Entwurf eines Mehrprozessor-Kommunikations-Controllers. Durch die parallele Verarbeitung von Protokollen soll die Leistung des Datenaustausches zwischen Rechensystemen erhöht werden, um der Übertragung von immer größer werdenden Datenmengen auf immer schnelleren Übertragungsmedien gerecht zu werden.

Petri-Netze bieten die Möglichkeit, parallele Prozesse formal zu beschreiben [Bau90]. Auf eine Beschreibung der Petri-Netze soll verzichtet werden, da sie den Umfang der vorliegenden Arbeit sprengen würde. Grundlegende Aussagen über Petri-Netze sind in [Rei82], [Son91] und [Pet81] zu finden.

Von der GMD Darmstadt wurden im Rahmen des Forschungsprojekts PROSIT² die *Produktnetze* als spezielle Form von höheren Petri-Netzen zur Spezifikation und Implementierung von Protokollen entwickelt [BOP89]. Diese weisen im Vergleich zu den „normalen“ Petri-Netzen die folgenden Charakteristika auf:

- Die Marken können einen *Wert* besitzen.
- Die Transitionen besitzen einen booleschen Ausdruck als *erweiterte Schaltbedingung*. Zum Schalten einer Transition ist nicht nur die Existenz von Marken in den vorliegenden Stellen erforderlich, sondern der boolesche Ausdruck muß ebenfalls erfüllt sein.
- Kantenbeschriftungen geben an, wie die Marken beim Schalten einer Transition verändert werden. Sie beeinflussen die Schaltregel einer Transition.

Auf der Basis der eingeschränkten Produktnetze [Rup91] sind im Verlauf des Forschungsprojekts PIKOM die PENCIL³-Netze entwickelt worden. Diese besitzen verschiedene Arten von Stellen (Einfach-, Mehrfach- und FIFO-Stellen). Als Schnittstelle zwischen dem als PENCIL-Netz formalisierten Kommunikationsprotokoll und der zu verwendenden Hardware wurde die Programmiersprache PENCIL/C⁴ entwickelt, die eine Erweiterung von ANSI-C ist. Eine genaue Definition von PENCIL/C ist in [Eng92] enthalten. Das Protokoll wird als PENCIL-Netz in PENCIL/C implementiert und durch einen speziellen Compiler [Gro93] in ANSI-C-Code übersetzt. Der so erhaltene Programm-Code kann dann auf verschiedenen Hardware-Plattformen, für die ein ANSI-C-Compiler existiert, zur Ausführung gebracht werden. Im augenblicklichen Stadium des Forschungsprojekts PIKOM ist noch keine Hardware-Realisierung der PAPER-Architektur auf der Basis eines Parallelrechners existent. Es wurde aber ein Emulator entwickelt, durch den

¹ Petri Net Based Parallel Architecture for Protocol Engineering and Realization

² PROtokoll Spezifikation, Implementierung und Test

³ PEtri Net based Communication Protocol Implementation Language

⁴ PEtri Net based Communication Protocol Implementation Language/extending C

im Zusammenwirken mit dem PENCIL/C-Compiler Aussagen über das Verhalten der PAPER-Architektur möglich sind. Desweiteren wird an einer Implementierung der PAPER-Architektur auf einem Transputer-Cluster gearbeitet ([Str94], [Pet94]).

3.1 Das funktionale Konzept der PAPER-Architektur

Das Ziel der PAPER-Architektur ist die parallele Verarbeitung von Protokollen durch einen Mehrprozessor-Kommunikations-Controller. Die Organisation der parallel arbeitenden Komponenten des Protokolls erfolgt dabei durch ein PENCIL-Netz. Die Transitionen des PENCIL-Netzes bilden die PENCIL/C-Prozeduren, durch die die Aktionen des Protokolls beschrieben und ausgeführt werden. Sie bestehen aus einer Aktivierungsbedingung und einem Anweisungsteil. Ist die Aktivierungsbedingung erfüllt, *kann* die Transition schalten. Aus Transitionssicht bilden die vor- und nachliegenden Stellen die Ein- und Ausgabeparameter der Prozeduren. Gemeinsame Stellen dienen den Transitionen zur Kommunikation untereinander. Das PENCIL-Netz beschreibt die Datenabhängigkeiten zwischen den einzelnen Protokoll-Prozeduren und steuert somit den Ablauf der Prozeduren.

Die Abarbeitung des als PENCIL-Netz formalisierten Protokolls auf einer Mehrprozessor-Hardware setzt innerhalb der PAPER-Architektur eine zweistufige Hierarchie voraus. Auf der *oberen* Hierarchiestufe müssen die Schaltbedingungen und Interferenzen der Transitionen kontrolliert werden, während die *untere* Stufe das Schalten (Feuern) der Transitionen übernimmt. Die PAPER-Architektur besteht daher aus einer *Kontrolleinheit* und einer *Ausführungseinheit*.

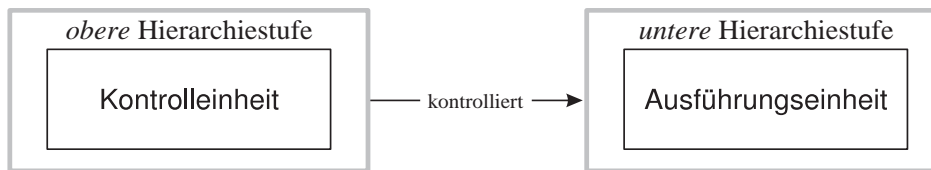


Abbildung 3.1: Das funktionale Konzept der PAPER-Architektur

3.2 Die Elemente der PAPER-Architektur

Wie oben beschrieben, setzt das parallele Abarbeiten eines PENCIL-Netzes innerhalb der PAPER-Architektur eine funktionale Trennung in zwei Ebenen voraus. Das ist zum einen die Kontrolleinheit, die die Ausführung des PENCIL-Netzes steuert und zum anderen die Ausführungseinheit, die die durch das PENCIL-Netz formalisierten Protokollaktionen ausführt. Die Kontrolleinheit stellt somit die Transitionen zur Verfügung, deren Transitionsfunktionen in der Ausführungseinheit abgearbeitet werden. Die Abbildung 3.2 zeigt den Aufbau und die Elemente der PAPER-Architektur, die im folgenden näher beschrieben werden.

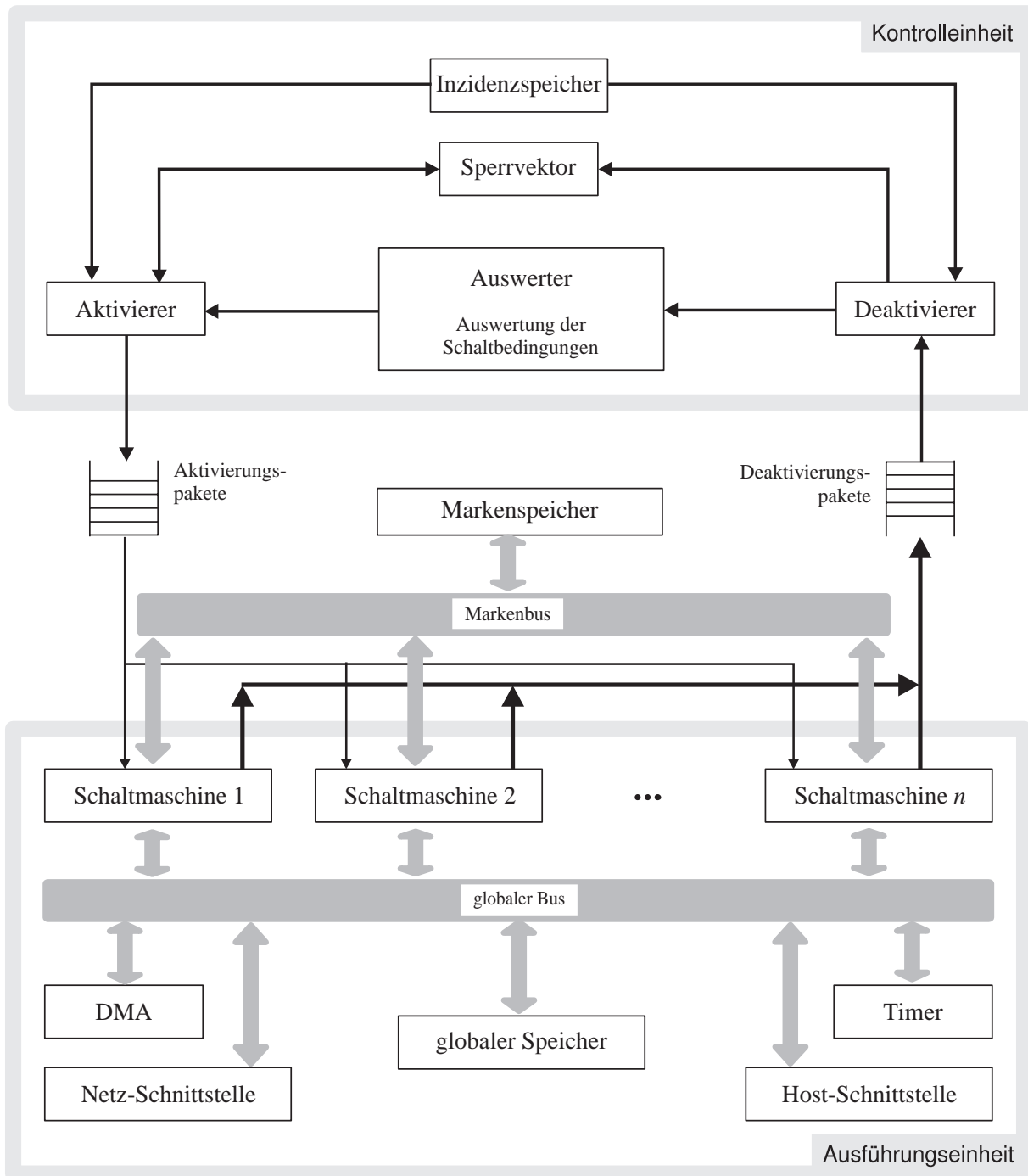


Abbildung 3.2: Die Elemente der PAPER-Architektur

3.2.1 Die Kontrolleinheit

Die Aufgabe der Kontrolleinheit ist es, aufgrund der aktuellen Markierung des PENCIL-Netzes schaltfähige Transitionen zu ermitteln. Die Leistungsfähigkeit der Kontrolleinheit bestimmt somit im wesentlichen auch die Leistungsfähigkeit der gesamten PAPER-Architektur. Je mehr schaltfähige Transitionen je Zeiteinheit gefunden werden, desto größer ist die Anzahl von möglicherweise parallel auszuführenden Protokollaktionen. Ist nämlich die Zeit für die Ermittlung von aktivierten Transitionen größer als deren Ausführungszeit, so wird die Möglichkeit der parallelen

Ausführung von Transitionsfunktionen eingeschränkt oder sogar unmöglich gemacht.

Bei der Ermittlung von schaltfähigen Transitionen muß die Konsistenz des PENCIL-Netzes sichergestellt sein. Die Möglichkeit des Schaltens einer Transition hängt sowohl von der Schaltbedingung als auch vom zugehörigen Eintrag im Sperrvektor ab. Eine Transition t ist gesperrt, wenn sie eine Eingabestelle hat, die gleichzeitig Ausgabestelle einer schaltenden Transition t' ist, welche den Inhalt dieser Stelle verändern kann.

Die Kontrolleinheit besteht im wesentlichen aus den drei ausführenden Komponenten *Auswerter*, *Aktivierer* und *Deaktivierer*. Diese Komponenten können aus einem Standardprozessor oder einem für die jeweilige Aufgabe zugeschnittenen Spezialprozessor bestehen. Durch eine *Pipeline*-Verarbeitung wird eine Parallelisierung bei der Transitionsaktivierung erreicht.

- Der **Auswerter** wertet aus der aktuellen Markierung des PENCIL-Netzes die Schaltbedingungen der Transitionen aus. Er übergibt einen Bitvektor an den Aktivierer, der anzeigt welche Transitionen schaltfähig sind.
- Der **Aktivierer** übernimmt die vom Auswerter als schaltfähig erkannten Transitionen. Diese Transitionen werden auf Interferenzen mit anderen, gerade im Schaltvorgang befindlichen Transitionen, getestet. Dazu werden die vom Auswerter als Bitvektor übergebenen Transitionen mit dem Inhalt des Sperrvektors verglichen. Ist eine als schaltfähig erkannte Transition dort nicht gesperrt, so wird sie aktiviert. Dabei sperrt der Aktivierer alle Transitionen, die durch gemeinsame Stellen im Nachliegerbereich der aktivierten Transition zu dieser indirekt inzident sind. Die zugehörigen Informationen beinhaltet der Inzidenzspeicher. Danach wird die aktivierte Transition zum Schalten an die Ausführungseinheit weitergeleitet.
- Der **Deaktivierer** nimmt die Transitionen, die in der Ausführungseinheit geschaltet haben von dieser entgegen. Er leitet die Marken, die durch das Schalten der Transition generiert worden sind, an den Auswerter weiter und entsperrt die Transitionen, die vom Aktivierer zum Schalten der zu bearbeitenden Transitionen gesperrt wurden.

Neben den ausführenden Komponenten besitzt die Kontrolleinheit noch zwei Speicherelemente:

1. Der **Inzidenzspeicher** enthält die statischen Informationen über die Struktur des PENCIL-Netzes, d.h. über die Inzidenzen dieses Netzes.
2. Der **Sperrvektor** beinhaltet die Informationen über die Verfügbarkeit von Transitionen. Für jede Transition existiert dort ein Eintrag welcher anzeigt, ob eine Transition gesperrt ist oder nicht. Durch den Sperrvektor werden die Zugriffe auf den Markenspeicher, der die aktuelle Markierung des PENCIL-Netzes enthält, synchronisiert. Dadurch ist gesichert, daß die Transitionen auf der aktuellen Markierung des PENCIL-Netzes arbeiten und somit keine unzulässigen Protokollaktionen ausgeführt werden können.

3.2.2 Die Ausführungseinheit

Die Ausführungseinheit besteht aus mehreren **Schaltmaschinen**. Darunter sind normale, sequenziell arbeitende Prozessoren zu verstehen, die das Schalten der von der Kontrolleinheit aktivierten Transitionen, also das Ausführen der Transitionsfunktion, übernehmen. Die Schaltmaschinenanzahl der PAPER-Architektur hängt von der Anzahl der verfügbaren Prozessoren der Parallelarchitektur ab und kann auf die Anzahl der parallel ausführbaren Protokollaktionen abgestimmt werden.

Die Schaltmaschinen sind über einen Bus mit dem **Markenspeicher** verbunden. Dieser enthält die aktuelle Markierung des gesamten PENCIL-Netzes. Durch den Markenspeicher wird der aktuelle Zustand des PENCIL-Netzes und somit auch des Protokolls wiedergegeben.

Neben dem Markenbus sind die Schaltmaschinen über einen weiteren Bus mit der **Host-** und der **Netzwerkschnittstelle** verbunden. Über letztere werden Datenpakete gesendet, bzw. empfangen. Ereignisse an der Host- oder Netzwerkschnittstelle erzeugen Interrupts, die von der Interrupt-Logik der Schaltmaschinen bearbeitet werden und das Erzeugen von Marken in bestimmten Stellen verursachen. Über diesen globalen Bus sind auch **Timer** mit den Schaltmaschinen verbunden. Diese werden unter bestimmten Umständen während des Schaltens von Transitionen gesetzt. Ist ein solcher Timer abgelaufen, wird ein Interrupt ausgelöst. Über den globalen Bus können die Schaltmaschinen noch mit weiteren Komponenten, wie z.B. DMA⁵-Bausteine, verbunden werden.

Zur Behandlung von Interrupts und anderen externen Ereignissen ist innerhalb der PAPER-Architektur kein explizites Ausführungselement vorgesehen. Deshalb müssen alle diese Ereignisse durch die Ausführungseinheit verwaltet werden. Dabei reicht es jedoch aus, nur auf das Auftreten eines solchen Ereignisses zu reagieren und dieses dann der Kontrolleinheit mitzuteilen. Auf diese Weise kann ein weiteres Reagieren auf externe Ereignisse direkt im PENCIL-Netz spezifiziert werden.

Die PAPER-Architektur stellt einen **globalen Speicher** zur Verfügung. Dieser kann von den Transitionen zur Speicherung von *gemeinsam* benötigten Daten benutzt werden. Jeder Zugriff durch die Schaltmaschinen auf den globalen Speicher erfolgt über den globalen Bus, weshalb von der Benutzung des globalen Speichers im Hinblick auf eine möglichst verzögerungsfreie Protokollverarbeitung nur wenig Gebrauch gemacht werden sollte.

3.2.3 Die Verbindung von Kontroll- und Ausführungseinheit

Die Kontroll- und die Ausführungseinheit sind durch zwei FIFO⁶-Warteschlangen miteinander verbunden (⇨ Abbildung 3.2). Eine Warteschlange verbindet den Aktivierer mit den Schaltmaschinen. Über diese Warteschlange gelangen die aktivierten Transitionen als Aktivierungspakete

⁵ **Direct Memory Access**: Diese Bausteine können selbständig Daten sowohl im Speicher umkopieren als auch von einer Schnittstelle in den Speicher und umgekehrt

⁶ **First In, First Out**

zum Schalten in die Schaltmaschinen. Die andere Warteschlange nimmt die Transitionen, die geschaltet haben, als Deaktivierungspakete von den Schaltmaschinen entgegen und leitet sie zum Deaktivierer weiter. Durch diese beiden Warteschlangen ist eine Synchronisation zwischen der Kontrolleinheit und der Ausführungseinheit gewährleistet.

Wie im Kapitel 2 ausgeführt wurde, erfolgt die Leistungsbewertung der PAPER-Architektur durch die Verwendung von stochastischen Bewertungsmethoden. In diesem Zusammenhang lassen sich durch Warteschlangenmodelle Bewertungsaussagen mit einem vertretbaren Aufwand machen. Der Vorteil der Warteschlangenmodelle liegt – im Gegensatz zu den stochastischen Petri-Netzen – in der einfachen Handhabbarkeit der zugrunde liegenden mathematischen Modelle. Die Elemente der Warteschlangentheorie, die für den weiteren Verlauf dieser Arbeit von Bedeutung sind, werden im folgenden Kapitel vorgestellt.

4. Warteschlangentheorie

Die Warteschlangentheorie hat Ihre Wurzeln in den Arbeiten von A.K. Erlang am Anfang dieses Jahrhunderts [Erl09]. Nach anfänglichen Akzeptanzschwierigkeiten entwickelte sich daraus im Laufe der Zeit eine eigene Wissenschaft. Seit den sechziger Jahren wird die Warteschlangentheorie als Hilfsmittel bei der Modellierung und Leistungsbewertung eingesetzt.

4.1 Die Beschreibung von Warteschlangensystemen

Innerhalb der Warteschlangentheorie versucht man Antworten auf Fragen der Art

- Wie lange muß ein Kunde warten?
- Wieviele Kunden befinden sich in der Warteschlange?
- Wie groß muß der Warteraum dimensioniert werden?

zu finden. Der Begriff „*Kunde*“ umfaßt ein breites Spektrum an Bedeutungen. Seine genaue Bedeutung ist vom jeweiligen Kontext abhängig.

Ein Warteschlangensystem setzt sich aus zwei Komponenten zusammen:

1. der *Warteschlange* (queue),
2. der *Bedienstation* (service center).

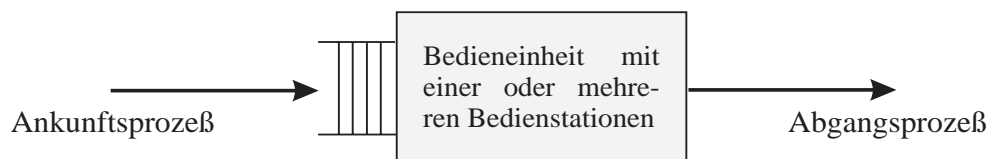


Abbildung 4.1: Der allgemeine Aufbau eines Warteschlangensystems

Die Kunden betreten das Warteschlangensystem über die Warteschlange. Sie verweilen dort, falls die Bedienstation durch einen anderen Kunden belegt ist, oder werden bei einer freien Bedienstation sofort bedient.

Ein Warteschlangensystem läßt sich durch die folgenden Charakteristika beschreiben:

- den *Ankunftsprozeß* der Kunden
- den *Bedienprozeß* innerhalb einer Bedienstation
- die Anzahl der *Bedieneinheiten* (server) innerhalb der Bedienstation
- die *Warteschlangendisziplin* (Bedienstrategie)

- die *Systemkapazität*

Der Ankunftsprozeß der Kunden

Der Ankunftsprozeß bildet den *Input* eines Warteschlangensystems. Dieser wird durch die durchschnittliche Anzahl von Kunden je Zeiteinheit (mittlere Ankunftsrate) oder durch die Zeit, die durchschnittlich zwischen aufeinanderfolgenden Ankünften liegt (mittlere Zwischenankunftszeit) quantifiziert. Sind diese Werte bekannt, spricht man von einem *deterministischen Ankunftsprozeß*. Lassen sich die Werte für die mittlere Ankunftsrate oder die mittlere Zwischenankunftszeit nicht mit Sicherheit angeben, so spricht man von einem *stochastischen Ankunftsprozeß*. Ein stochastischer Ankunftsprozeß wird durch eine Wahrscheinlichkeitsverteilung und den zugehörigen stochastischen Prozeß beschrieben.

Der Ankunftsprozeß läßt sich auch durch sein Zeitverhalten charakterisieren. Verändern sich die Parameter der Wahrscheinlichkeitsverteilung im Laufe der Zeit, so nennt man den Ankunftsprozeß *nicht-stationär*. Ist er zeitunabhängig, so wird er *stationär* genannt.

Der Bedienprozeß innerhalb der Bedienstation

Die Beschreibung des Bedienprozesses ist ähnlich dem des Ankunftsprozesses. So wird auch der Bedienprozeß durch eine Rate (Anzahl von bedienten Kunden je Zeiteinheit) oder durch die Zeit (Bedienzeit für einen Kunden) quantifiziert. Ebenso kann der Bedienprozeß deterministisch oder stochastisch sein. Im letzten Fall benötigt man eine Wahrscheinlichkeitsverteilung zur Beschreibung des Bedienprozesses.

Die Anzahl der Bedieneinheiten innerhalb der Bedienstation

Die Bedienstation kann aus einer oder mehreren identischen parallel arbeitenden Bedieneinheiten bestehen.

Die Warteschlangendisziplin (Bedienstrategie)

Durch die Bedienstrategie wird festgelegt, welcher Kunde aus der Warteschlange als nächstes bedient wird. Die wichtigsten Bedienstrategien sind:

FCFS (First-Come-First-Served):

Die Kunden werden in der Reihenfolge ihrer Ankunft bedient.

LCFS (Last-Come-First-Served):

Der zuletzt angekommene Kunde wird als nächster bedient.

SIRO (Service-In-Random-Order):

Die Auswahl des nächsten Kunden erfolgt zufällig.

RR (Round Robin):

Ist die Bedienung eines Kunden nach einer fest vorgegebenen Zeitspanne noch nicht beendet, so wird der Kunde verdrängt und wieder in die Warteschlange eingereiht, die nach *FCFS* abgearbeitet wird. Dies wird so oft wiederholt, bis der Kunde vollständig bedient worden ist.

PS (Processor Sharing):

Entspricht *RR* mit sehr kleinen Zeitspannen. Dadurch entsteht der Eindruck, als ob alle Kunden gleichzeitig bedient würden mit entsprechend längerer Bedienzeit.

Statische Prioritäten:

Die Auswahl erfolgt nach fest vorgegebenen Prioritäten der Kunden.

Dynamische Prioritäten:

Die Auswahl erfolgt nach dynamischen Prioritäten, die sich in Abhängigkeit der Zeit ändern.

Prioritäts-Disziplinen können *unterbrechend* (preemptive) oder *nichtunterbrechend* (nonpreemptive) sein.

Unterbrechende Prioritäts-Disziplin:

Ein gerade in Bedienung befindlicher Kunde niedriger Priorität wird bei der Ankunft eines Kunden höherer Priorität unterbrochen. Nach der Abarbeitung des höher priorisierten Kunden wird die Bedienung des Kunden mit der niedrigeren Priorität fortgesetzt.

Nichtunterbrechende Prioritäts-Disziplin:

Erst nach Abarbeitung des gerade in Bedienung befindlichen Kunden wird ein Kunde einer höheren Prioritätsklasse bedient.

Ist die Kundenanzahl innerhalb einer Prioritätsklasse größer Eins, so muß innerhalb der Prioritätsklasse eine Auswahl getroffen werden.

Die Systemkapazität

Die Systemkapazität bezieht sich auf die physikalische Größe der Warteschlange. Wenn die Warteschlangenlänge diese Größe erreicht hat, können neue Kunden das System erst wieder betreten, wenn ein Kunde fertig bedient worden ist.

4.1.1 Kendall'sche Notation

Zur einheitlichen Beschreibung von Wartesystemen hat sich die *Kendall'sche Notation* [Ken53] durchgesetzt. Diese besteht aus einer Folge von Buchstaben und Zahlen, die durch / getrennt werden. Die generische Kendall-Notation lautet:

$$A/B/c/n/p/S$$

Hierbei kennzeichnet A die Verteilung der Zwischenankunftszeiten und B die Verteilung der Bedienzeiten. Die Anzahl der identischen Bedieneinheiten innerhalb der Bedienstation wird mit c ($c \geq 1$) angegeben. Das vierte Feld, n , gibt die Systemkapazität an. Durch den Buchstaben p wird die Anzahl von potentiellen Kunden quantifiziert. Die letzte Position der generischen Kendall Notation bezieht sich auf die Warteschlangendisziplin.

Für A (die Verteilung der Zwischenankunftszeit) und B (die Verteilung der Bedieneinheiten) werden traditionell die folgenden Symbole verwendet:

M	Exponentialverteilung ¹ ,
E_k	Erlang-Verteilung mit k Phasen,
H_k	Hyperexponentialverteilung mit k Phasen,
C_k	Cox-Verteilung mit k Phasen,
D	Deterministische Verteilung, d.h. die Zwischenankunfts- bzw. Bedienzeiten sind konstant,
G	Allgemeine Verteilung,
GI	Allgemeine unabhängige Verteilung.

Oft werden zur Beschreibung von Wartesystemen nur die ersten drei Positionen der Kendall'schen Notation angegeben. Dann geht man von einer unendlichen Systemkapazität, einer unendlichen Anzahl von potentiellen Kunden und der *FCFS*-Warteschlangendisziplin aus.

4.2 Wahrscheinlichkeitstheoretische Grundlagen

Wie schon im vorangegangenen Abschnitt angedeutet, wird in der Warteschlangentheorie das Verhalten eines Modells durch einen stochastischen Prozeß modelliert. Dementsprechend sind auch die Leistungsmerkmale stochastisch verteilt.

Da zur Analyse von Modellen, die auf stochastischen Warteschlangensystemen basieren, bestimmte Kenntnisse der Wahrscheinlichkeitstheorie Voraussetzung sind, soll im folgenden ein einleitender Überblick über diese gegeben werden.

4.2.1 Zufallsvariablen

Eine *Zufallsvariable* X ist eine Funktion, die das Ereignis eines zufallsbedingten Vorganges ausdrückt. In einem stochastischen Warteschlangensystem ist z.B. die Zeit zwischen der Ankunft zweier aufeinanderfolgender Kunden oder die Verweilzeit der Kunden im Warteschlangensystem eine Zufallsvariable. Zufallsvariablen können kontinuierliche oder diskrete Werte annehmen. Aus diesem Grund unterscheidet man zwischen *kontinuierlichen* und *diskreten* Zufallsvariablen.

¹ Die Exponentialverteilung wird durch M bezeichnet, um Konflikte mit der Erlang-Verteilung zu vermeiden. Es wurde M gewählt, da die Exponentialverteilung die Markov-Eigenschaft erfüllt.

4.2.1.1 Diskrete Zufallsvariablen

Eine Zufallsvariable χ , die nur diskrete Werte annehmen kann, wird als *diskrete Zufallsvariable* bezeichnet. Die diskreten Werte sind im allgemeinen ganze Zahlen. Die diskrete Zufallsvariable wird durch die möglichen Werte, die sie annehmen kann und die Wahrscheinlichkeiten für diese Werte beschrieben. Die Menge dieser Wahrscheinlichkeiten nennt man *Wahrscheinlichkeitsverteilung*.

Die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen χ ist durch die *Wahrscheinlichkeitsfunktion*

$$p_k = P\{\chi = k\}, \quad k = 0, 1, 2, \dots \quad (4.1)$$

gegeben. Also durch die Wahrscheinlichkeiten, daß die Zufallsvariable χ den Wert k annimmt. Dabei muß gelten:

$$\begin{aligned} P\{\chi = k\} &\geq 0, & \forall k \\ \sum_{\text{alle } k} P\{\chi = k\} &= 1 \end{aligned}$$

Aus der Wahrscheinlichkeitsfunktion einer diskreten Verteilung (\Leftrightarrow Gleichung (4.1)) lassen sich weitere wichtige Parameter ableiten:

- *Erwartungswert* oder *Mittelwert*:

$$\bar{\chi} = E[\chi] = \sum_{\text{alle } k} k \cdot P\{\chi = k\} \quad (4.2)$$

- *n-tes Moment*:

$$\bar{\chi}^n = E[\chi^n] = \sum_{\text{alle } k} k^n \cdot P\{\chi = k\}$$

Das n -te Moment ist der Erwartungswert der n -ten Potenz von χ . Das erste Moment von χ ist der Erwartungswert von χ .

- *n-tes zentrales Moment*:

$$\overline{(\chi \Leftrightarrow \bar{\chi})^n} = E[(\chi \Leftrightarrow E[\chi])^n] = \sum_{\text{alle } k} (k \Leftrightarrow \bar{\chi})^n \cdot P\{\chi = k\}$$

Das n -te zentrale Moment ist der Erwartungswert der n -ten Potenz der Differenz zwischen χ und dem Erwartungswert von χ .

- Das zweite zentrale Moment bezeichnet man als die *Varianz* von χ :

$$\sigma_{\chi}^2 = \text{var}(\chi) = \overline{(\chi \Leftrightarrow \bar{\chi})^2} = \bar{\chi}^2 \Leftrightarrow \bar{\chi}^2$$

σ_{χ} heißt *Standardabweichung*.

- Der *Varianzkoeffizient* ist die normierte Standardabweichung:

$$c_{\chi} = \frac{\sigma_{\chi}}{\bar{\chi}}.$$

Der Varianzkoeffizient c_{χ} , die Standardabweichung σ_{χ} und die Varianz $\text{var}(\chi)$ geben an, wie

stark der Wert der Zufallsvariablen χ im Mittel vom Erwartungswert abweicht. Ist

$$c_\chi = \sigma_\chi = \text{var}(\chi) = 0,$$

so bedeutet das, daß die Zufallsvariable mit Wahrscheinlichkeit Eins einen festen Wert einnimmt.

4.2.1.2 Kontinuierliche Zufallsvariablen

Kann eine Zufallsvariable alle reellen Werte eines Intervalls $[a, b] \subseteq \mathbb{R}$ annehmen, d.h. $-\infty < a < b < +\infty$, wird sie als *kontinuierliche Zufallsvariable* bezeichnet. Eine kontinuierliche Zufallsvariable läßt sich durch die zugehörige *Verteilungsfunktion*

$$F_\chi(x) = P\{\chi \leq x\}$$

beschreiben. Diese gibt für alle Werte x aus der Wertemenge von χ die Wahrscheinlichkeit dafür an, daß der Wert der Zufallsvariablen χ kleiner oder gleich einem betrachteten x ist.

Statt der Verteilungsfunktion kann die *Dichtefunktion* verwendet werden, die die erste Ableitung der Verteilungsfunktion ist.

$$f_\chi(x) = \frac{dF_\chi(x)}{dx}$$

Die Dichtefunktion einer kontinuierlichen Zufallsvariablen entspricht der Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen. Daher erhält man die Formeln für den Erwartungswert und die höheren Momente einer kontinuierlichen Zufallsvariablen aus denen für diskrete Zufallsvariablen, indem man die Wahrscheinlichkeitsfunktion durch die Dichtefunktion und die Summation durch die Integration ersetzt.

Zwei weitere wichtige Eigenschaften der Dichtefunktion sind:

$$f_\chi(x) \geq 0, \quad \forall x$$

$$\int_{-\infty}^{\infty} f_\chi(x) dx = 1 .$$

4.2.1.3 Zufallsvariablen in einem Warteschlangensystem

Wie in Abbildung 4.2 dargestellt, gibt es bei der stochastischen Analyse von Warteschlangensystemen eine Reihe von Zufallsvariablen.

- τ *Zwischenankunftszeit*: Die Zeit zwischen zwei aufeinanderfolgenden Ankünften von Kunden.
- q *Wartezeit* (queueing time): Das Zeitintervall zwischen der Ankunftszeit und dem Zeitpunkt, an dem mit der Bedienung des Kunden begonnen wird.
- s *Bedienzeit* (service time) für einen Kunden.
- r *Verweilzeit* (response time): Als Verweilzeit wird die Gesamtheit der Zeit bezeichnet, die ein Kunde im Wartesystem verbringt. Sie ist gleich der Summe aus der Wartezeit

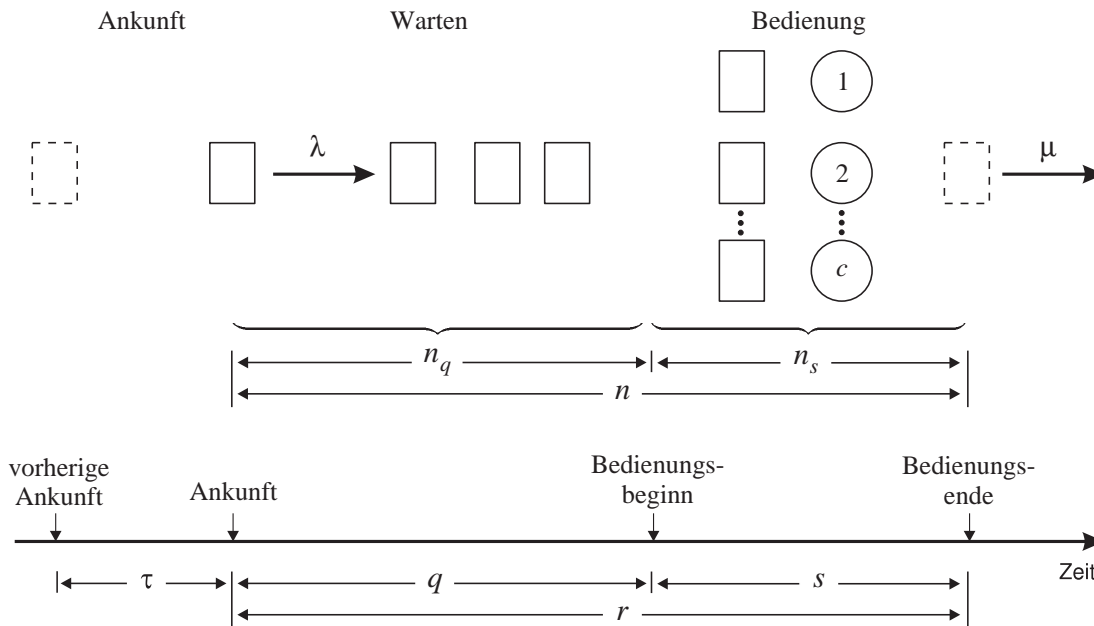


Abbildung 4.2: Zufallsvariablen in einem Warteschlangensystem

q und der Bedienzeit s

$$r = q + s$$

Da r , w und s Zufallsvariablen sind, gilt für ihre Mittelwerte im stationären Zustand:

$$E[r] = E[q] + E[s]. \quad (4.3)$$

$E[r]$ ist die *mittlere Verweilzeit* eines Kunden im Wartesystem.

n_q Anzahl von Kunden, die in der Warteschlange auf Bedienung warten (Warteschlangenlänge).

n_s Anzahl von Kunden, die bedient werden.

n Anzahl von Kunden im Warteschlangensystem. Die Anzahl der Kunden in einem Warteschlangensystem ist gleich der Summe aus der Anzahl der wartenden Kunden und der Anzahl in Bedienung befindlichen Kunden.

$$n = n_q + n_s$$

Da n , n_q und n_s Zufallsvariablen sind, gilt für ihre Mittelwerte im stationären Zustand:

$$E[n] = E[n_q] + E[n_s]. \quad (4.4)$$

Neben den oben genannten Zufallsvariablen gibt es innerhalb eines Warteschlangensystems noch zwei weitere wichtige Parameter:

λ *Ankunftsrate*

Das ist die mittlere Anzahl von Kunden, die pro Zeiteinheit eintreffen. Die Ankunftsrate ist der Kehrwert der mittleren Zwischenankunftszeit $\bar{\tau}$.

μ *Bedienrate*

Durch die Bedienrate wird die mittlere Anzahl von Kunden, die je Zeiteinheit bedient werden, quantifiziert. Die gesamte Bedienrate für eine Bedienstation mit c Bedieneinheiten ist $c \cdot \mu$

4.2.2 Wahrscheinlichkeitsverteilungen

Bei der analytischen Leistungsbewertung mittels Warteschlangenmodellen sind die Zwischenankunftszeit und die Bedienzeit eines Warteschlangensystems Zufallsgrößen und durch die zugehörige *Wahrscheinlichkeitsverteilung* festgelegt.

Die in der Warteschlangentheorie gebräuchlichsten Wahrscheinlichkeitsverteilungen sind die

- Poisson-Verteilung,
- Exponentialverteilung,
- Erlang-k-Verteilung,
- k-Phasen-Hyperexponentialverteilung.

Da im Rahmen dieser Arbeit nur von der *Poisson-Verteilung* und der *Exponentialverteilung* Gebrauch gemacht wird, sollen diese nun vorgestellt werden.

Definition 4.1: (*Poisson-Verteilung*)

Sind beliebige Ereignisse voneinander unabhängig und gibt die diskrete Zufallsvariable X die Anzahl dieser Ereignisse an, dann ist X *poissonverteilt* mit dem Parameter λ .

$$P\{X=k\} = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

Der Parameter $\lambda > 0$ gibt die Rate an, mit der die Ereignisse eintreffen, d.h. im Mittel treten pro Zeiteinheit λ Ereignisse ein. Der Erwartungswert und die Varianz einer poissonverteilten Zufallsvariablen X sind gleich:

$$E[X] = \text{var}(X) = \lambda.$$

Die Poisson-Verteilung wird bei der Modellierung einer Vielzahl von zufallsbedingten Vorgängen, bei denen die einzelnen Ereignisse unabhängig sind, benutzt:

- Anzahl von ankommenden Kunden in einem Warteschlangensystem,
- Anzahl von Anrufen in einem Fernsprechnet,
- Anzahl von Tippfehlern je Seite,
- Anzahl von Geburts- und Sterbefällen in einer Stadt.

Neben der Poisson-Verteilung ist die Exponentialverteilung in der Warteschlangentheorie die wichtigste Wahrscheinlichkeitsverteilung.

Definition 4.2: (*Exponentialverteilung*)

Eine kontinuierliche Zufallsvariable χ heißt *exponentiell verteilt* mit dem Parameter λ , falls ihre Dichte gegeben ist durch die Funktion

$$f_{\chi}(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & , \quad x > 0 \\ 0 & , \quad x \leq 0 \end{cases} .$$

Die Verteilungsfunktion ist dann gegeben durch

$$F_{\chi}(x) = P\{\chi \leq x\} = \begin{cases} 1 - e^{-\lambda \cdot x} & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases} .$$

Der Erwartungswert ist $E[\chi] = \frac{1}{\lambda}$ und die Varianz ist $\text{var}(\chi) = \frac{1}{\lambda^2}$.

Die Bedeutung der Exponentialverteilung liegt darin, daß sie die einzige kontinuierliche Wahrscheinlichkeitsverteilung ist, die die *Markov-Eigenschaft der Gedächtnislosigkeit* (memoryless property) besitzt. Ist die Zwischenankunftszeit τ in einem Warteschlangensystem exponentiell verteilt mit im Mittel $\frac{1}{\lambda}$, dann ist die Verteilungsfunktion gegeben durch:

$$F_{\tau}(t) = P\{\tau \leq t\} = 1 - e^{-\lambda \cdot t}, \quad t \geq 0.$$

Die Eigenschaft der Gedächtnislosigkeit läßt sich durch die Gleichung

$$P\{\tau \leq s + t \mid \tau > s\} = P\{\tau \leq t\} \tag{4.5}$$

beschreiben. Ist die Zufallsvariable τ die Wartezeit auf ein bestimmtes Ereignis, dann hat die Gleichung (4.5) die folgende Bedeutung [Lan92]: Wurde bereits s Zeiteinheiten auf das Ereignis gewartet, dann ist die Wahrscheinlichkeit dafür höchstens weitere t Zeiteinheiten zu warten unabhängig von s . Es existiert keine Erinnerung daran, wie lange schon gewartet wurde.

$$\begin{aligned} P\{\tau \leq s + t \mid \tau > s\} &= \frac{P\{s \leq \tau \leq s + t\}}{P\{\tau > s\}} \\ &= \frac{P\{\tau \leq s + t\} - P\{\tau \leq s\}}{P\{\tau > s\}} \\ &= \frac{(1 - e^{-\lambda \cdot (s+t)}) - (1 - e^{-\lambda \cdot s})}{1 - (1 - e^{-\lambda \cdot s})} \\ &= 1 - e^{-\lambda \cdot t} \\ &= P\{\tau \leq t\} \end{aligned}$$

Die mittlere Zeit bis zur Ankunft des nächsten Kunden ist also immer $\frac{1}{\lambda}$, unabhängig von der Zeit seit dem letzten Ankunftszeitpunkt.

Ein weiteres wichtiges Charakteristikum der Exponentialverteilung ist die sogenannte *Poisson-Eigenschaft*. Sind die Zwischenankunftszeit oder Bedienzeit in einem Warteschlangensystem exponentiell verteilt, dann ist die Zufallsvariable für die Anzahl der Kunden, die in einem festen Zeitintervall eintreffen bzw. bedient werden poissonverteilt. Man spricht dann auch von einem *Poisson'schen Ankunfts- bzw. Bedienprozeß* (\Leftrightarrow Abschnitt 4.2.3.4).

4.2.3 Stochastische Prozesse

Bei der analytischen Modellbildung und Leistungsbewertung wird das dynamische Ablaufgeschehen mit Hilfe von Zufallsvariablen beschrieben. Wenn das Warteschlangenmodell zum Zeitpunkt t durch die Zufallsvariable $\chi(t)$ beschrieben wird, so läßt es sich zum Zeitpunkt t' durch eine andere Zufallsvariable $\chi(t')$ beschreiben. Das Verhalten des Warteschlangensystems ist somit eine Menge von Zufallsvariablen $\{\chi(t), t \in \mathcal{T}\}$, bei der sich die einzelnen Zufallsgrößen durch den Zeitparameter t voneinander unterscheiden.

Definition 4.3: (*Stochastischer Prozeß*)

Eine Familie von Zufallsvariablen $\{\chi(t), t \in \mathcal{T}\}$ heißt *stochastischer Prozeß*. Dabei ist \mathcal{T} die *Indexmenge* des Prozesses. Die Menge aller Werte, die die Zufallsvariable $\chi(t)$ annehmen kann, heißt *Zustandsraum* \mathcal{Z} des stochastischen Prozesses. Jeder einzelne Wert $\chi(t)$ wird *Zustand* des Prozesses zur Zeit t genannt.

4.2.3.1 Klassifizierung stochastischer Prozesse

Wird der Zustandsraum \mathcal{Z} aus einem oder mehreren Intervallen der reellen Zahlen gebildet, so heißt der stochastische Prozeß *zustandskontinuierlich*. Ist der Zustandsraum dagegen endlich oder abzählbar, so ist der stochastische Prozeß *zustandsdiskret*. Zustandsdiskrete Prozesse werden auch als *Ketten* bezeichnet.

Stochastische Prozesse lassen sich auch bezüglich der Indexmenge \mathcal{T} klassifizieren. Ist \mathcal{T} eine endliche oder abzählbare Menge von Zeitpunkten, so spricht man von einem *zeitdiskreten* stochastischen Prozeß. Zeitdiskrete Prozesse werden auch als *Folgen* bezeichnet. Ist \mathcal{T} ein endliches oder unendliches Zeitintervall, dann wird der stochastische Prozeß als *zeitkontinuierlich* bezeichnet.

Das Hauptkriterium zur Klassifizierung von stochastischen Prozessen ist die *Abhängigkeit* der Zufallsvariablen voneinander. Es wird hierbei zwischen *Markov-Prozessen* und *Nicht-Markov-Prozessen* unterschieden.

Ein stochastischer Prozeß wird *Markov-Prozeß* genannt, wenn er die Markov-Eigenschaft besitzt. Das ist dann der Fall, wenn die zukünftigen Zustände unabhängig von den Zuständen der Vergangenheit sind. Der zukünftige Verlauf des stochastischen Prozesses ist nur vom augenblicklichen Zustand $\chi(t)$ abhängig und nicht von den vorherigen Zuständen. Besitzt ein stochastischer Prozeß diese Eigenschaft nicht, so wird er der Klasse der Nicht-Markov-Prozesse zugeordnet.

Für die Leistungsbewertung von Rechensystemen und deren Komponenten sind vor allem die zustandsdiskreten Markov-Prozesse (Markov-Ketten) von zentraler Bedeutung.

4.2.3.2 Markov-Ketten

Definition 4.4: (Markov-Kette)

Ein stochastischer Prozeß $\{\chi(t), t \in \mathcal{T}\}$ ist eine *zeitkontinuierliche Markov-Kette*, wenn für alle $n \in \mathbb{N}$ und $x_k \in \mathcal{Z}$ und alle $\{t_1, t_2, \dots, t_{n+1}\}$ mit $t_1 < t_2 < \dots < t_{n+1}$ gilt:

$$\begin{aligned} P\{\chi(t_{n+1}) = x_{n+1} \mid \chi(t_1) = x_1, \chi(t_2) = x_2, \dots, \chi(t_n) = x_n\} = \\ P\{\chi(t_{n+1}) = x_{n+1} \mid \chi(t_n) = x_n\} \end{aligned} \quad (4.6)$$

Durch die Gleichung (4.6) wird die Markov-Eigenschaft zum Ausdruck gebracht. Die Wahrscheinlichkeitsverteilung der Zeit, in der sich der Prozeß in den einzelnen Zuständen aufhält, hat somit die Eigenschaft der Gedächtnislosigkeit. Die Exponentialverteilung ist die einzige kontinuierliche Wahrscheinlichkeitsverteilung, die diese Eigenschaft besitzt. Daraus ergibt sich die Folgerung, daß alle Zustandszeiten einer zeitkontinuierlichen Markov-Kette *exponentiell verteilt* sein müssen.

Der Ausdruck der rechten Seite in Gleichung (4.6) gibt die *Übergangswahrscheinlichkeiten* der Markov-Kette an. Dafür wird die folgende Schreibweise eingeführt:

$$p_{ij}(s, t) = P\{\chi(t) = j \mid \chi(s) = i\}, \quad t > s. \quad (4.7)$$

$p_{ij}(s, t)$ ist also die bedingte Wahrscheinlichkeit dafür, daß sich der stochastische Prozeß zum Zeitpunkt t im Zustand j befindet, falls er zum Zeitpunkt $s < t$ im Zustand i ist.

Sind die Übergangswahrscheinlichkeiten unabhängig vom Zeitpunkt t und hängen nur von der Zeitdifferenz $\tau = t \Leftrightarrow s$ ab, so wird die Markov-Kette *homogen* genannt. Die Gleichung für die Übergangswahrscheinlichkeiten (4.7) kann bei homogenen Markov-Ketten vereinfacht geschrieben werden:

$$p_{ij}(\tau) = P\{\chi(s + \tau) = j \mid \chi(s) = i\}.$$

Für die Übergangswahrscheinlichkeiten einer homogenen Markov-Kette gilt:

$$\begin{aligned} p_{ij}(\tau) &\geq 0 && \forall i, j \in \mathcal{Z} \\ \sum_{j \in \mathcal{Z}} p_{ij}(\tau) &= 1 && \forall i \in \mathcal{Z} \\ p_{ij}(\tau) &= \sum_{k \in \mathcal{Z}} p_{ik}(\tau \Leftrightarrow \theta) \cdot p_{kj}(\theta) \end{aligned} \quad (4.8)$$

Die Gleichung (4.8) ist unter dem Namen *Chapman-Kolmogorov-Gleichung* für homogene Markov-Ketten bekannt.

Nicht nur die Übergangswahrscheinlichkeiten p_{ij} sind das Ziel bei der Untersuchung von Markov-Prozessen, sondern vielmehr die Bestimmung der *Zustandswahrscheinlichkeiten*

$$p_j(t) = P\{\chi(t) = j\}. \quad (4.9)$$

Die Zustandswahrscheinlichkeit $p_j(t)$ gibt die Wahrscheinlichkeit an, mit der sich der Prozeß zum Zeitpunkt t im Zustand j befindet. Diese Zustandswahrscheinlichkeiten lassen sich unter

Zuhilfenahme der allgemeinen Anfangsbedingungen $p_i(0) = P\{\chi(0) = i\}$ durch

$$p_j(t) = \sum_{i \in \mathcal{Z}} p_{ij}(t) \cdot p_i(0), \quad j \in \mathcal{Z}$$

ermitteln. Die Markov-Kette ist somit durch die Initialwahrscheinlichkeiten $p_i(0)$ und durch die Übergangswahrscheinlichkeiten vollständig definiert.

Die Bestimmung der Übergangswahrscheinlichkeiten unter Benutzung der Chapman-Kolmogorov-Gleichung (\Leftrightarrow Gleichung (4.8)) ist jedoch schwierig. Diese werden daher mit Hilfe der *Übergangsraten* des Prozesses vom Zustand i in den Zustand j ermittelt.

$$q_{ij} = \lim_{t \rightarrow 0} \frac{p_{ij}(t) \Leftrightarrow \delta_{ij}}{t} \quad \text{mit} \quad \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & \text{sonst} \end{cases}$$

Da $\Leftrightarrow q_{ii}$ die Rate ist, mit der der Zustand i verlassen wird, gilt:

$$\sum_{j \in \mathcal{Z}} q_{ij} = 0, \quad \forall i \in \mathcal{Z}.$$

Durch Differenzieren der Gleichung (4.9) nach t und durch Benutzen sowohl der Kolmogorov-Rückwärtsgleichung

$$\frac{dp_{ij}(t)}{dt} = \sum_{k \in \mathcal{Z}} p_{ik}(t) \cdot q_{kj},$$

als auch der Kolmogorov-Vorwärtsgleichung

$$\frac{dp_{ij}(t)}{dt} = \sum_{k \in \mathcal{Z}} p_{kj}(t) \cdot q_{ik}$$

kann man die zeitabhängigen Übergangswahrscheinlichkeiten des Prozesses durch Lösen des Differentialgleichungssystems

$$\frac{dp_j(t)}{dt} = \sum_{i \in \mathcal{Z}} p_i(t) \cdot q_{ij} \quad (4.10a)$$

bestimmen. Dieses Gleichungssystem läßt sich mit dem Zeilenvektor $\vec{p}(t)$ und der *Übergangsratenmatrix* Q durch

$$\frac{d\vec{p}(t)}{dt} = \vec{p}(t) \cdot Q. \quad (4.10b)$$

beschreiben. Die Diagonalelemente der Übergangsratenmatrix Q

$$Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} & \cdots \\ q_{21} & q_{22} & q_{32} & \cdots \\ q_{31} & q_{32} & q_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

ergeben sich aus $q_{ii} = \sum_{j \neq i} q_{ij}$.

Ein einfaches Hilfsmittel zur Beschreibung von Markov-Ketten sind die *Zustandstransitionsdiagramme*. Darunter versteht man gerichtete Graphen, deren Knoten die Zustände und deren Kanten die möglichen Übergänge zwischen den Zuständen der Markov-Kette repräsentieren. Die Kanten sind mit den Übergangsraten beschriftet.

Bei der Ermittlung der zeitabhängigen Zustandswahrscheinlichkeiten $\vec{p}(t)$ durch das Gleichungssystem (4.10) erhält man oftmals gar keine oder nur sehr komplexe Lösungen. Aus diesem Grunde wird das Hauptinteresse bei der Untersuchung von Markov-Ketten auf *zeitlose* Zustände (*stationäre* Zustände) gelegt.

Ein *stationärer Zustand* wird erreicht, wenn der stochastische Prozeß hinreichend lange Zeit gelaufen ist, so daß der Einfluß des Anfangszustands vernachlässigbar wird. Der stochastische Prozeß befindet sich dann im *statistischen Gleichgewicht* und die Zustandswahrscheinlichkeiten werden *stationäre Zustandswahrscheinlichkeiten* oder auch *Gleichgewichtszustandswahrscheinlichkeiten* genannt.

Damit die Markov-Kette einen stationären Zustand erreichen kann, müssen ihre Zustände verschiedene Bedingungen erfüllen.

Definition 4.5: (*Irreduzibilität*)

Eine Markov-Kette $\{X(t), t \in T\}$ heißt *irreduzibel*, falls jeder Zustand j von jedem anderen Zustand i erreicht werden kann.

Definition 4.6: (*Rekurrenz*)

Ein Zustand einer Markov-Kette heißt *rekurrent*, wenn der Prozeß mit einer Wahrscheinlichkeit von Eins wieder in diesen Zustand zurückkehrt. Ist die Rückkehrzeit endlich, wird der Zustand *positiv-rekurrent* genannt. Bei einer unendlichen Rückkehrzeit spricht man von einem *null-rekurrenten* Zustand.

Definition 4.7: (*Ergodizität*)

Eine irreduzible Markov-Kette heißt *ergodisch*, wenn sämtliche Zustände positiv-rekurrent sind.

Für eine irreduzible homogene Markov-Kette existieren stets die Grenzwahrscheinlichkeiten $\{p_j, j \in \mathcal{Z}\}$ mit

$$p_j = \lim_{t \rightarrow \infty} p_{ij}(t), \quad \forall j \in \mathcal{Z}.$$

Die Grenzwahrscheinlichkeiten einer ergodischen Markov-Kette heißen *stationäre Wahrscheinlichkeiten* und erfüllen die Beziehungen:

$$\begin{aligned} p_j &= \lim_{t \rightarrow \infty} p_j(t), & \forall j \in \mathcal{Z}, \\ \lim_{t \rightarrow \infty} \frac{dp_j(t)}{dt} &= 0, & \forall j \in \mathcal{Z}. \end{aligned} \tag{4.11}$$

Unter Benutzung des Gleichungssystems (4.10) und unter Hinzunahme der Normalisierungsbedingung

$$\sum_{i \in \mathcal{Z}} p_i = 1 \tag{4.12}$$

erhält man eine eindeutige Lösung für die stationären Zustandswahrscheinlichkeiten durch Lösen des homogenen Gleichungssystems

$$\sum_{i \in \mathcal{Z}} = p_i \cdot q_{ij} = 0 \tag{4.13}$$

Auf der Basis der Markov-Prozesse existieren zwei weitere stochastische Prozesse, die für die Warteschlangentheorie von Bedeutung sind:

1. Geburts-/Sterbeprozesse
2. Poisson-Prozesse

4.2.3.3 Geburts-/Sterbeprozesse

Geburts-/Sterbeprozesse sind eine spezielle Form der Markov-Ketten. Sie besitzen die einschränkende Bedingung, daß Übergänge ausschließlich zwischen zwei benachbarten Prozeßzuständen möglich sind.

Definition 4.8: (*Geburts-/Sterbeprozess*)

Ein *Geburts-/Sterbeprozess* ist eine zeitkontinuierliche Markov-Kette über den Zustandsraum $\mathcal{Z} = \{0, 1, 2, \dots\}$, bei der nur Zustandsübergänge der Form ± 1 erlaubt sind.

Die Abbildung 4.3 zeigt das Zustandstransitionsdiagramm eines Geburts-/Sterbeprozesses.

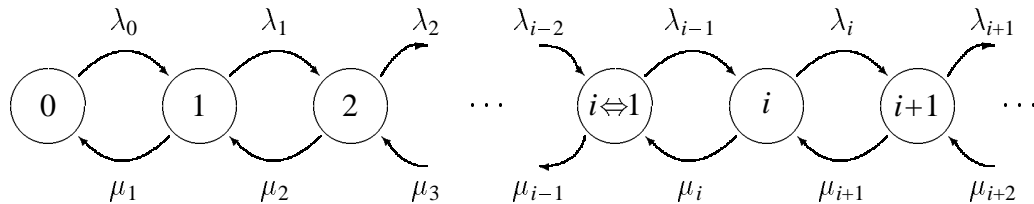


Abbildung 4.3: Zustandstransitionsdiagramm eines Geburts-/Sterbeprozesses

Die Parameter λ_i und μ_i sind die *Geburts-* und *Sterberaten* des Zustands i . Sie repräsentieren die Rate, mit der der stochastische Prozeß vom Zustand i in den Zustand $i + 1$ bzw. $i \Leftrightarrow 1$ übergeht. Die Elemente der Übergangsratenmatrix Q ergeben sich somit zu:

$$q_{ij} = \begin{cases} \lambda_i & , \quad j = i + 1 \\ \mu_i & , \quad j = i \Leftrightarrow 1 \\ \Leftrightarrow(\lambda_i + \mu_i) & , \quad i = j \\ 0 & , \quad \text{sonst} \end{cases} .$$

Ferner gilt für die Geburtsraten

$$\lambda_i > 0, \quad \forall i$$

und die Sterberaten

$$\mu_0 = 0 \quad \text{und} \quad \mu_i > 0 \quad \forall i > 0.$$

Als Untermenge der Markov-Ketten sind auch bei den Geburts-/Sterbeprozessen die Zustandswahrscheinlichkeiten $p_n(t) = P\{X(t) = n\}$ von Interesse. Das sind die Wahrscheinlichkeiten, daß sich der Geburts-/Sterbeprozess zum Zeitpunkt t im Zustand n befindet. Die Zustandswahrscheinlichkeiten lassen sich durch ein aus Differentialgleichungen bestehendes Gleichungssystem nach

(4.10) nur sehr aufwendig ermitteln.

$$\begin{aligned}\frac{dp_i(t)}{dt} &= \Leftrightarrow (\lambda_i + \mu_i) \cdot p_i(t) + \lambda_{i-1} \cdot p_{i-1}(t) + \mu_{i+1} \cdot p_{i+1}(t), & \forall i \geq 1 \\ \frac{dp_0(t)}{dt} &= \Leftrightarrow \lambda_0 \cdot p_0(t) + \mu_1 \cdot p_1(t)\end{aligned}\quad (4.14)$$

Eine einfachere Berechnung der Zustandswahrscheinlichkeiten ist möglich, wenn sich der Geburts-/Sterbeprozess in einem stationären Zustand befindet. In [Kle75] wird gezeigt, daß das genau dann der Fall ist, wenn:

$$\exists k_0, \forall k > k_0 \quad \frac{\lambda_k}{\mu_k} < 1. \quad (4.15)$$

Ist das der Fall, dann ist der Geburts-/Sterbeprozess ergodisch und es existieren die Grenzwerte aus Gleichung (4.11). Das Gleichungssystem (4.13) zur Ermittlung der stationären Zustandswahrscheinlichkeiten wird unter der Annahme, daß $\mu_0 = \lambda_{-1} = 0$ durch

$$\begin{aligned}0 &= \Leftrightarrow (\lambda_i + \mu_i) \cdot p_i + \lambda_{i-1} \cdot p_{i-1} + \mu_{i+1} \cdot p_{i+1}, & \forall i \geq 1 \\ 0 &= \Leftrightarrow \lambda_0 \cdot p_0 + \mu_1 \cdot p_1\end{aligned}\quad (4.16)$$

gebildet.

Diese Gleichungen lassen sich direkt aus dem Zustandstransitionsdiagramm der Geburts-/Sterbeprozesse (\diamond Abbildung 4.3) herleiten.

Eine rekursive Lösung des Gleichungssystems (4.16) ist durch

$$p_{i+1} = \frac{\lambda_i}{\mu_{i+1}} \cdot p_i, \quad \mu_{i+1} \neq 0, \forall i \in \mathcal{Z}$$

möglich. Durch fortgesetztes Einsetzen erhält man für die Berechnung der stationären Zustandswahrscheinlichkeiten eines ergodischen Geburts-/Sterbeprozesses:

$$p_i = p_0 \cdot \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}}, \quad \forall i \in \mathcal{Z}. \quad (4.17a)$$

Die Initialwahrscheinlichkeit p_0 läßt sich mittels der Normalisierungsbedingung für Wahrscheinlichkeiten $\sum_{i \in \mathcal{Z}} p_i = 1$ (\diamond Gleichung (4.12)) durch Einsetzen der Gleichung (4.17a) aufstellen:

$$\begin{aligned}1 &= \sum_{i \in \mathcal{Z}} p_i = \sum_{i \in \mathcal{Z}} p_0 \cdot \left(\prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}} \right) \\ &= p_0 \cdot \sum_{i \in \mathcal{Z}} \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}} \\ &= p_0 \cdot \left(1 + \sum_{\substack{i \in \mathcal{Z} \\ i \neq 0}} \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}} \right).\end{aligned}$$

Dies ist äquivalent zu:

$$p_0 = \left[1 + \sum_{\substack{i \in \mathcal{Z} \\ i \neq 0}} \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}} \right]^{-1}. \quad (4.17b)$$

Insgesamt lassen sich die stationären Zustandswahrscheinlichkeiten eines ergodischen Geburts-/Sterbeprozesses mit den Gleichungen (4.17a) und (4.17b) durch

$$p_i = \frac{\prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}}}{1 + \sum_{\substack{i \in \mathcal{Z} \\ i \neq 0}} \prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}}}, \quad \forall i \in \mathcal{Z}$$

berechnen.

Der letzte stochastische Prozeß, der im Rahmen der wahrscheinlichkeitstheoretischen Grundlagen für den Einsatz von Warteschlangensystemen bei der analytischen Leistungsbewertung vorgestellt werden soll ist der Poisson-Prozeß.

4.2.3.4 Der Poisson-Prozeß

Der *Poisson-Prozeß* ist ein *reiner* Geburtsprozeß mit einer konstanten Rate $\lambda > 0$. Er kann als ein Spezialfall der Geburts-/Sterbeprozesse angesehen werden. Dabei sind für alle Zustände $i \in \mathcal{Z}$ die Sterberaten $\mu_i = 0$ und die Geburtsraten $\lambda_i = \lambda, \forall i \geq 0$. Der Poisson-Prozeß läßt somit nur Zustandsänderungen der Größe +1 zu. Es ist ihm nicht möglich, einen Zustand zweimal zu erreichen. Aus diesem Grunde existieren *keine* stationären Zustände.

Im Falle des Poisson-Prozesses können die Differentialgleichungen (4.10) bzw. (4.16) vereinfacht und direkt gelöst werden.

$$\frac{dp_i(t)}{dt} = \Leftrightarrow \lambda \cdot p_i(t) + \lambda \cdot p_{i-1}(t), \quad \forall i \geq 1 \quad (4.18a)$$

$$\frac{dp_0(t)}{dt} = \Leftrightarrow \lambda \cdot p_0(t) \quad (4.18b)$$

Zur Vereinfachung sei angenommen, daß der Prozeß zum Zeitpunkt $t = 0$ im Zustand 0 beginnt.

$$p_i(0) = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0 \end{cases}$$

Die Differentialgleichung (4.18b) ergibt dann:

$$p_0(t) = e^{-\lambda \cdot t} \quad (4.19)$$

Alle anderen Gleichungen können jetzt rekursiv gelöst werden und man erhält für den allgemeinen Fall:

$$p_i(t) = \frac{(\lambda \cdot t)^i}{i!} \cdot e^{-\lambda \cdot t}, \quad \forall i \in \mathcal{Z}.$$

Die Zustandswahrscheinlichkeiten erzeugen also eine Poisson-Verteilung mit dem Parameter $\lambda \cdot t$. Eine Zusammenfassung der hier vorgestellten stochastischen Prozesse wird in Abbildung 4.4 gegeben.

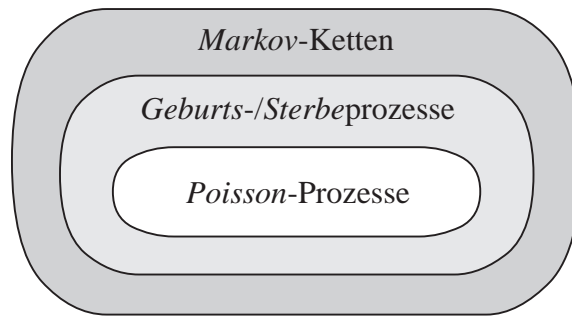


Abbildung 4.4: Der Zusammenhang verschiedener stochastischer Prozesse

4.2.3.5 Eigenschaften des Poisson-Prozesses

Dem Poisson-Prozeß kommt im Rahmen der Warteschlangentheorie eine zentrale Bedeutung zu. Er besitzt zum einen einige einfache Eigenschaften und zum anderen gibt es eine Vielzahl von physikalischen oder organischen Prozessen, die sich als Poisson-Prozeß modellieren lassen.

Wie im Abschnitt 4.2.3.4 hergeleitet, gilt der folgende Satz:

Satz 4.9:

Sei $\{\chi(t), t \in \mathcal{T}\}$ ein Poisson-Prozeß mit Rate $\lambda > 0$. Dann ist die Zufallsvariable χ , welche die Anzahl der Ereignisse in einem Intervall der Länge $t > 0$ beschreibt, poissonverteilt mit dem Parameter $\lambda \cdot t$.

$$P\{\chi = k\} = \frac{(\lambda \cdot t)^k}{k!} \cdot e^{-\lambda \cdot t}$$

Die mittlere Anzahl der Ereignisse in einem Intervall der Länge t ist nach Abschnitt 4.2.2 $E[\chi] = \lambda \cdot t$. Die mittlere Anzahl der Ereignisse pro Zeiteinheit ist gleich der Rate des Poisson-Prozesses. In Bezug auf die Warteschlangentheorie läßt sich das Auftreten von Ereignissen als Folge von Ankünften, die mit einer Ankunftsrate λ eintreffen, interpretieren.

Zwischen dem Poisson-Prozeß und der im Abschnitt 4.2.2 vorgestellten Exponentialverteilung besteht eine sehr enge Verbindung [Gro74]:

Es sei \tilde{t} eine Zufallsvariable, die die Zeit zwischen zwei aufeinanderfolgenden Ankünften angibt. Dann gilt für die Verteilungsfunktion, daß die Zeit zwischen zwei Ankünften $\leq t$ ist:

$$A(t) = P\{\tilde{t} \leq t\} = 1 \Leftrightarrow P\{\tilde{t} > t\}.$$

$P\{\tilde{t} > t\}$ ist aber genau die Wahrscheinlichkeit, daß im Zeitintervall $(0, t)$ keine Ankunft erfolgt. Das ist $p_0(t)$. Mit Gleichung (4.19) erhält man die Wahrscheinlichkeitsverteilungsfunktion

$$A(t) = 1 \Leftrightarrow e^{-\lambda \cdot t}, \quad t \geq 0,$$

deren Dichtefunktion

$$a(t) = \frac{dA(t)}{dt} = \lambda \cdot e^{-\lambda \cdot t}, \quad t \geq 0$$

ist. Die beiden letzten Gleichungen sind die Verteilungs- und die Dichtefunktion der Exponentialverteilung. Der gerade hergeleitete Zusammenhang zwischen einem Poisson-Prozeß und der

Exponentialverteilung läßt sich in dem folgenden Satz ausdrücken.

Satz 4.10:

Sei $\{\chi(t), t \in \mathcal{T}\}$ ein Poisson-Prozeß mit Rate λ . Sei weiter $0 < t_1 < t_2 < \dots$ die Folge der Zeitpunkte, mit der Ankünfte erfolgen und sei $\tau_1 = t_1, \tau_2 = t_2 \Leftrightarrow t_1, \dots, \tau_k = t_k \Leftrightarrow t_{k-1}, \dots$ die Folge der Zwischenankunftszeiten. Dann sind die τ_n identisch unabhängig exponentiell verteilt mit dem Erwartungswert $E[\tau_n] = \frac{1}{\lambda}$.

Somit hat ein Poisson'scher Ankunftsprozeß exponentiell verteilte Zwischenankunftszeiten.

Der Poisson-Prozeß hat für die Warteschlangentheorie zwei weitere wichtige Eigenschaften [Bol89]:

1. Faßt man n Poisson-Prozesse mit ihren Zwischenankunftszeitverteilungen $1 \Leftrightarrow e^{-\lambda_i t}, 1 \leq i \leq n$, zu einem einzigen Prozeß zusammen, so erhält man wieder einen Poisson-Prozeß, der die Zwischenankunftszeit $1 \Leftrightarrow e^{-\lambda t}, \lambda = \sum_{i=1}^n \lambda_i$, besitzt.

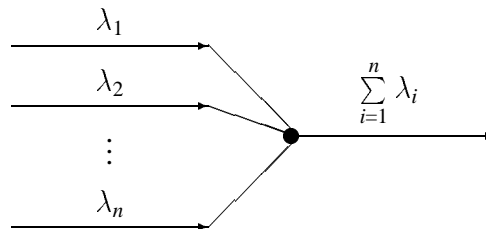


Abbildung 4.5: Verschmelzen von Poisson-Prozessen

2. Wird ein Poisson-Prozeß mit einer Zwischenankunftszeitverteilung von $1 \Leftrightarrow e^{-\lambda t}$ so in n Prozesse aufgespalten, daß die Ankünfte mit Wahrscheinlichkeit $p_i, 1 \leq i \leq n$, beim i -ten Prozeß ankommen, dann besitzt der i -te Teilprozeß die Zwischenankunftszeitverteilung $1 \Leftrightarrow e^{-p_i \lambda t}$, d.h. es entstehen n Poisson-Prozesse.

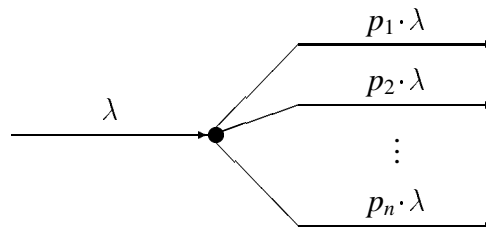


Abbildung 4.6: Aufspalten eines Poisson-Prozesses

4.3 Elementare Wartesysteme

Nachdem im vorangegangenen Abschnitt wichtige wahrscheinlichkeitstheoretische Grundlagen für die Warteschlangentheorie eingeführt worden sind, werden in diesem Abschnitt einige Wartesysteme und deren Leistungsgrößen beschrieben.

Die Vorstellung von Wartesystemen beschränkt sich auf die Typen von Wartesystemen, die im weiteren Verlauf dieser Arbeit von Bedeutung sind. Das sind Wartesysteme, bei denen sowohl der Ankunfts- als auch der Bedienprozeß der Markov-Eigenschaft genügen, d.h. die Zwischenankunftszeit und Bedienzeit sind exponentiell verteilt.

Wie schon im Abschnitt 4.1 beschrieben, besteht ein Warteschlangensystem aus einer *Warteschlange* und einer *Bedienstation*, die aus einer oder mehreren identischen *Bedieneinheiten* besteht (\Leftrightarrow Abbildung 4.1). Für ein solches Warteschlangensystem werden die Bezeichnungen „Wartesystem“, „elementares Wartesystem“, „Bediensystem“ oder „Knoten“ verwendet.

4.3.1 Leistungsgrößen eines Wartesystems

Die Ermittlung von *Leistungsgrößen* ist das primäre Ziel der Modellierung eines beliebigen Systems durch Wartesysteme. Mit Hilfe der Leistungsgrößen können dann Aussagen über die Leistungsfähigkeit des Systems gemacht werden.

Die Leistungsgrößen sind, da sie das dynamischen Ablaufgeschehen innerhalb des Wartesystems ausdrücken, zeitabhängig. Im Abschnitt 4.2.3 wurde darauf hingewiesen, daß das Ermitteln von zeitabhängigen Zustandswahrscheinlichkeiten eine schwierige, oftmals erfolglose Aufgabe ist. So interessiert man sich bei der Ermittlung der Leistungsgrößen in der Regel für Ergebnisse in einem *stationären Systemzustand*. In einem solchen Zustand sind alle Auswirkungen des Initialzustands abgeklungen und die Leistungsgrößen zeitunabhängig. Das Wartesystem befindet sich im *statistischen Gleichgewicht*. Im statistischen Gleichgewicht ist die Rate, mit der Kunden im Wartesystem ankommen, gleich der Rate, mit der sie das Wartesystem wieder verlassen. Damit sich ein Wartesystem im statistischen Gleichgewicht befindet, müssen bestimmte Voraussetzungen erfüllt sein (\Leftrightarrow Gleichung (4.21)).

Es folgt eine Beschreibung der Basisleistungsgrößen für Wartesysteme.

Zustandswahrscheinlichkeit p_k :

Das Systemverhalten eines Wartesystems kann in vielen Fällen mit den Zustandswahrscheinlichkeiten p_k ausreichend beschrieben werden. Hieraus lassen sich die Mittelwerte aller anderen Leistungsgrößen ableiten. Es gilt:

$$p_k = P\{\text{es befinden sich } k \text{ Aufträge im Wartesystem}\}.$$

Auslastung ρ :

Die Auslastung ρ gibt den Bruchteil der Gesamtzeit an, den die Bedienstation aktiv ist. In einem Wartesystem mit c Bedieneinheiten wird die Auslastung durch

$$\rho = \frac{\lambda}{c \cdot \mu}, \quad c \geq 1, \quad (4.20)$$

berechnet.

Durch die Auslastung ρ läßt sich die Bedingung für die Existenz eines statistischen Gleich-

gewichts formulieren. Es muß die Bedingung

$$\rho < 1 \quad (4.21)$$

erfüllt sein. Pro Zeiteinheit dürfen im Mittel nicht mehr Aufträge ankommen als bedient werden können. Diese Gleichgewichtsbedingung gilt für *alle* Aussagen in dieser Arbeit.

Weitere Leistungsgrößen sind die im Abschnitt 4.2.1.3 beschriebenen Zufallsvariablen für Wartesysteme.

4.3.2 Das Gesetz von Little

Eine grundlegende Aussage im Rahmen der Warteschlangentheorie ist das „Gesetz von Little“ [Lit61]. Es stellt einen Zusammenhang zwischen der *mittleren Gesamtanzahl* von Kunden und deren *mittleren Verweilzeit* im Wartesystem her.

Sei $\{\alpha(t), t \geq 0\}$ ein stochastischer Prozeß, der die Anzahl der Kunden, die innerhalb des Zeitintervalls $[0, t]$ an einem Wartesystem ankommen, zählt.

$$\alpha(t) = \text{Ankünfte der Kunden im Zeitintervall } [0, t]$$

Ein weiterer stochastischer Prozeß $\{\delta(t), t \geq 0\}$ zählt die Kunden, die innerhalb des Zeitintervalls $[0, t]$ das Wartesystem verlassen.

$$\delta(t) = \text{Abgänge der Kunden im Zeitintervall } [0, t]$$

Dann ist die Anzahl der Kunden, die bis zum Zeitpunkt t im Wartesystem eingetroffen sind, aber dieses noch nicht verlassen haben:

$$N(t) = \alpha(t) \Leftrightarrow \delta(t).$$

$N(t)$ ist die Anzahl der Kunden im Wartesystem zum Zeitpunkt t .

Die gesamte Fläche zwischen den Funktionsgeraden (\Leftrightarrow Abbildung 4.7) der stochastischen Prozesse $\{\alpha(t), t \geq 0\}$ und $\{\delta(t), t \geq 0\}$ bis zu einem beliebigen Zeitpunkt τ ist die Gesamtzeit, die alle Kunden innerhalb des Zeitintervalls $[0, \tau]$ im Wartesystem verbracht haben. Diese Zeit sei mit $\gamma(t)$ bezeichnet.

$\lambda(\tau)$ wird als die durchschnittliche Ankunftsrate der Kunden im Zeitintervall $[0, \tau]$ definiert:

$$\lambda(\tau) = \frac{\alpha(\tau)}{\tau}. \quad (4.22a)$$

Sei nun $R(\tau)$ die durchschnittliche Verweilzeit eines Kunden im Wartesystems. Da $\gamma(\tau)$ als die Gesamtverweilzeit aller Kunden im Zeitintervall $[0, \tau]$ definiert ist, gilt für $R(\tau)$:

$$R(\tau) = \frac{\gamma(\tau)}{\alpha(\tau)}. \quad (4.22b)$$

Als letztes wird $N(\tau)$ als die durchschnittliche Kundenanzahl innerhalb des Wartesystems für das Zeitintervall $[0, \tau]$ definiert:

$$N(\tau) = \frac{\gamma(\tau)}{\tau}. \quad (4.22c)$$

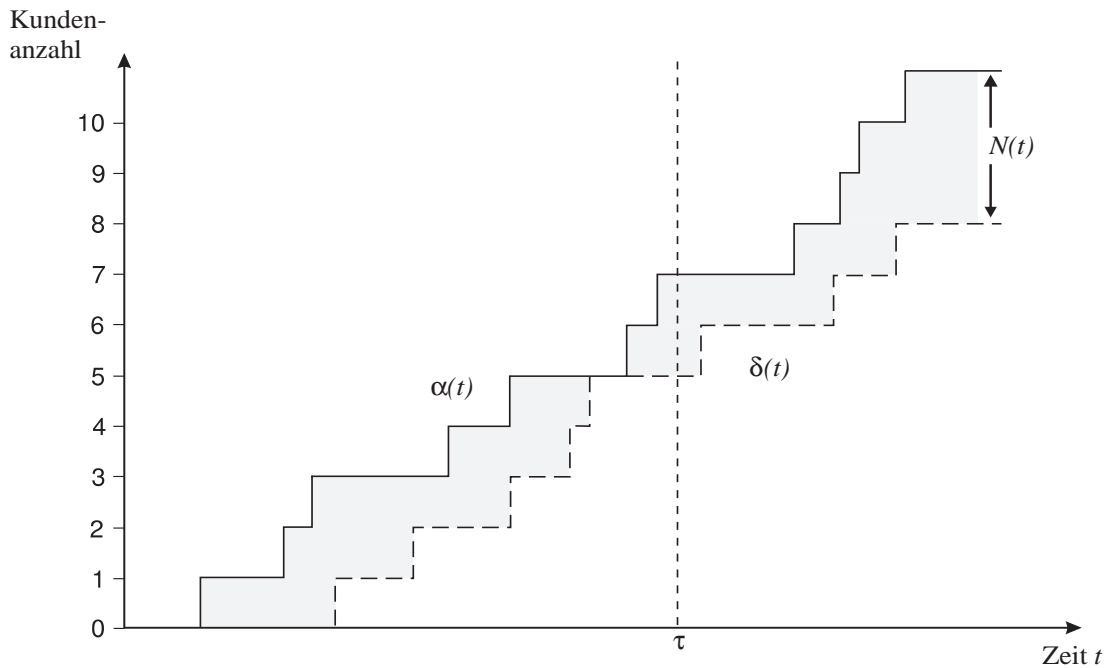


Abbildung 4.7: Die Verteilungsfunktionen der stochastischen Prozesse $\alpha(t)$ und $\delta(t)$

Durch Einsetzen der umgeformten Gleichungen (4.22a) und (4.22b) in die Gleichung (4.22c) erhält man:

$$\begin{aligned} N(\tau) &= \frac{\alpha(\tau) \cdot R(\tau)}{\tau} \\ &= \frac{\lambda(\tau) \cdot \tau \cdot R(\tau)}{\tau} \\ &= \lambda(\tau) \cdot R(\tau). \end{aligned}$$

Wenn sich das Wartesystem in einem stationären Zustand befindet dann existieren die Grenzwerte $\lambda = \lim_{\tau \rightarrow \infty} \lambda(\tau)$ und $R = \lim_{\tau \rightarrow \infty} R(\tau)$ (ϕ Abschnitt 4.2.3.2). Wenn diese Grenzwerte existieren, existiert auch N als Grenzwert für $N(t)$ und es gilt:

$$N = \lambda \cdot R.$$

Mit den Bezeichnungen für die Zufallsvariablen eines Wartesystems aus Abschnitt 4.2.1.3 ergibt sich dann:

$$E[n] = \lambda \cdot E[r]. \quad (4.23a)$$

Die Gleichung (4.23a) ist bekannt als das *Gesetz von Little* und stellt einen unmittelbaren Zusammenhang zwischen den Erwartungswerten für die *Anzahl von Kunden* und deren *Verweilzeit* im Wartesystem her.

Die oben hergeleiteten Beziehungen sind unabhängig vom Ankunfts- oder Bedienprozeß. Auch wird die Bedienstation und deren Verhalten nicht näher spezifiziert. Daher ist es auch möglich, aus der Gleichung (4.23a) eine Beziehung zwischen dem Erwartungswert der *Warteschlangenlänge* und der *Wartezeit* eines Kunden aufzustellen:

$$E[n_q] = \lambda \cdot E[q]. \quad (4.23b)$$

4.3.3 M/M/1-Wartesysteme

Das M/M/1-Wartesystem ist der gebräuchlichste Typ eines Warteschlangenmodells. Die Zwischenankunftszeit und die Bedienzeit sind exponentiell verteilt. Der Ankunftsprozeß ist ein Poisson-Prozeß mit der Ankunftsrate $\lambda > 0$. Gemäß der Kendall'schen Notation (\Leftrightarrow Abschnitt 4.1.1) besitzt ein M/M/1-Wartesystem

- eine Bedienstation mit nur einer Bedieneinheit,
- eine unbeschränkte Systemkapazität,
- eine unbeschränkte Anzahl von potentiellen Kunden,
- die Warteschlangendisziplin *FCFS*.

Zur Analyse eines Warteschlangenmodells vom Typ M/M/1 genügt die Kenntnis der Ankunftsrate λ und der Bedienrate μ . Der Zustand eines M/M/1-Wartesystems ist durch die Anzahl der Kunden innerhalb des M/M/1-Wartesystems gegeben, d.h. der Zustandsraum $\mathcal{Z} = \{0, 1, 2, \dots, \infty\}$ ist unendlich. Ein M/M/1-Wartesystem ist somit als ein Spezialfall der Geburts-/Sterbeprozesse (\Leftrightarrow Abschnitt 4.2.3.3) anzusehen, wobei für die Geburts- und Sterberaten gilt:

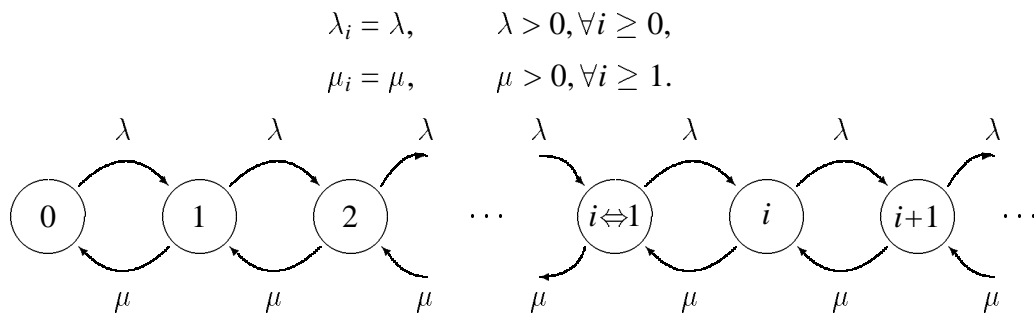


Abbildung 4.8: Zustandstransitionsdiagramm eines M/M/1-Wartesystems

Die Abbildung 4.8 zeigt das Zustandstransitionsdiagramm eines M/M/1-Wartesystems. Damit sich ein M/M/1-Wartesystem im statistischen Gleichgewicht befindet, muß gemäß Gleichung (4.15) und (4.20) für die Auslastung ρ gelten:

$$\rho = \frac{\lambda}{\mu} < 1 \quad (4.24)$$

Die stationären Zustandswahrscheinlichkeiten eines M/M/1-Wartesystems sind die Wahrscheinlichkeiten p_i , daß sich i Kunden im System befinden. Für diese gilt nach Gleichung (4.17a):

$$p_i = p_0 \cdot \prod_{k=0}^{i-1} \left(\frac{\lambda}{\mu} \right) \stackrel{(4.24)}{=} p_0 \cdot \prod_{k=0}^{i-1} \rho = p_0 \cdot \rho^i, \quad \forall i \in \mathcal{Z}. \quad (4.25a)$$

Die Initialwahrscheinlichkeit p_0 läßt sich durch die Normalisierungsbedingung für Wahrscheinlichkeiten ($\sum_{i \in \mathcal{Z}} p_i = 1$) und durch Einsetzen der Gleichung (4.25a) bestimmen durch:

$$1 = \sum_{i=0}^{\infty} p_i \stackrel{(4.25a)}{=} \sum_{i=0}^{\infty} p_0 \cdot \rho^i = p_0 \cdot \sum_{i=0}^{\infty} \rho^i \Leftrightarrow$$

$$p_0 = \left[\sum_{i=0}^{\infty} \rho^i \right]^{-1}$$

Da $|\rho| < 1$ ist, wird durch $\sum_{i=0}^{\infty} \rho^i$ die *geometrische Reihe* gebildet, die bekanntlich gegen $\frac{1}{1-\rho}$ konvergiert.

Für die Initialwahrscheinlichkeit p_0 eines M/M/1-Wartesystems gilt also:

$$p_0 = 1 \Leftrightarrow \rho. \quad (4.25b)$$

Insgesamt lassen sich die stationären Zustandswahrscheinlichkeiten eines M/M/1-Wartesystems aus den Gleichungen (4.25a) und (4.25b) berechnen:

$$p_i = p_0 \cdot \rho^i = (1 \Leftrightarrow \rho) \cdot \rho^i, \quad \forall i \in \mathcal{Z}. \quad (4.25c)$$

Aus Gleichung (4.25c) wird ersichtlich, daß die Anzahl der Kunden *geometrisch* verteilt ist. Die stationären Zustandswahrscheinlichkeiten bilden die Grundlage zur Berechnung aller weiteren Leistungsgrößen.

Die erste dieser Leistungsgrößen ist die *mittlere Kundenanzahl* innerhalb des M/M/1-Wartesystems $E[n]$:

$$E[n] \stackrel{(4.2)}{=} \sum_{i=0}^{\infty} i \cdot p_i \stackrel{(4.25c)}{=} \sum_{i=0}^{\infty} i \cdot (1 \Leftrightarrow \rho) \cdot \rho^i = (1 \Leftrightarrow \rho) \cdot \sum_{i=1}^{\infty} i \cdot \rho \quad (4.26a)$$

In [Gro74] wird gezeigt, daß sich obige Gleichung nach

$$E[n] = \frac{\rho}{1 \Leftrightarrow \rho} = \frac{\lambda}{\mu \Leftrightarrow \lambda}. \quad (4.26b)$$

auflösen läßt.

Mit dem Gesetz von Little läßt sich die *mittlere Verweilzeit* eines Kunden ermitteln:

$$E[r] \stackrel{(4.23a)}{=} \frac{E[n]}{\lambda} = \frac{1}{\mu \Leftrightarrow \lambda}. \quad (4.26c)$$

Die mittlere Verweilzeit $E[r]$ ist gleich der Summe aus der mittleren Wartezeit $E[q]$ und der mittleren Bedienzeit $E[s]$. Da die Bedienzeit exponentiell verteilt mit Rate μ ist, gilt für den Erwartungswert der *mittleren Bedienzeit*:

$$E[s] = \frac{1}{\mu}. \quad (4.26d)$$

Aus den Gleichungen (4.26c) und (4.26d) läßt sich die *mittlere Wartezeit* $E[q]$ ermitteln:

$$E[q] = E[r] \Leftrightarrow E[s] = \frac{1}{\mu \Leftrightarrow \lambda} \Leftrightarrow \frac{1}{\mu} = \frac{\lambda}{\mu \cdot (\mu \Leftrightarrow \lambda)}. \quad (4.26e)$$

Ebenfalls mit dem Gesetz von Little läßt sich, unter Benutzung des Ergebnisses für die mittlere Wartezeit, die *mittlere Warteschlangenlänge* $E[n_q]$ berechnen:

$$E[n_q] \stackrel{(4.23b)}{=} \lambda \cdot E[q] = \frac{\lambda^2}{\mu \cdot (\mu \Leftrightarrow \lambda)}. \quad (4.26f)$$

Die mittlere Kundenanzahl in einem M/M/1-Wartesystem ist gleich der Summe aus der mittleren Anzahl von Kunden in der Warteschlange $E[n_q]$ und der *mittleren Anzahl von Kunden in Bedienung* $E[n_s]$.

$$E[n] = E[n_q] + E[n_s]$$

Durch Umformen der letzten Gleichung und unter Benutzung von schon ermittelten Werten, kann die *mittlere Anzahl* von Kunden, die *bedient* werden, berechnet werden durch:

$$E[n_s] = E[n] \Leftrightarrow E[n_q] = \frac{\lambda}{\mu \Leftrightarrow \lambda} \Leftrightarrow \frac{\lambda^2}{\mu \cdot (\mu \Leftrightarrow \lambda)} = \frac{\lambda}{\mu} = \rho. \quad (4.26g)$$

Eine weitere interessante Größe bei der Analyse von M/M/1-Wartesystemen ist die Wahrscheinlichkeit, daß sich *mindestens* k Kunden im Wartesystem befinden. Sei n die Zufallsvariable, welche die Kundenanzahl eines M/M/1-Wartesystems angibt, dann ist:

$$\begin{aligned} P\{n \geq k\} &= \sum_{i=k}^{\infty} p_i \stackrel{(4.25c)}{=} \sum_{i=k}^{\infty} (1 \Leftrightarrow \rho) \cdot \rho^i \\ &= (1 \Leftrightarrow \rho) \cdot \rho^k \cdot \sum_{i=k}^{\infty} \rho^{i-k} \\ &= (1 \Leftrightarrow \rho) \cdot \rho^k \cdot \sum_{k'=0}^{\infty} \rho^{k'}. \end{aligned}$$

Die Summe $\sum_{k'=0}^{\infty} \rho^{k'}$ ist, da $|\rho| < 1$, eine geometrische Reihe mit dem Grenzwert $\frac{1}{1-\rho}$. Somit läßt sich die Wahrscheinlichkeit, daß sich mindestens k Kunden im Wartesystem befinden, berechnen durch:

$$P\{n \geq k\} = (1 \Leftrightarrow \rho) \cdot \rho^k \cdot \frac{1}{1 \Leftrightarrow \rho} = \rho^k. \quad (4.26h)$$

Oftmals ist bei der Analyse eines M/M/1-Wartesystems auch die *Wahrscheinlichkeit für eine bestimmte Kundenanzahl in der Warteschlange* interessant. Diese ist nach [Jai91]:

$$P\{n_q = k\} = P\{n = k + 1\} = \begin{cases} 1 \Leftrightarrow \rho^2 & , k = 0 \\ (1 \Leftrightarrow \rho) \cdot \rho^{k+1} & , k > 0 \end{cases} \quad (4.26i)$$

Die Tabelle A.1 im Anhang gibt einen Überblick über die in diesem Abschnitt hergeleiteten Formeln für die Analyse von M/M/1-Wartesystemen.

Der in diesem Abschnitt vorgestellte Markov'sche Warteschlangentyp M/M/1 ist oftmals nicht anwendbar, da das zu modellierende System *mehrere* identische parallel arbeitende Bedieneinheiten besitzt. Das Warteschlangenmodell, welches in einem solchen Fall zum Einsatz kommt, ist das M/M/c-Wartesystem.

4.3.4 M/M/c-Wartesysteme

Ein M/M/c-Wartesystem hat exponentiell verteilte Zwischenkunftszeiten und Bedienzeiten. Der Ankunftsprozeß ist ein Poisson-Prozeß mit der Ankunftsrate $\lambda > 0$. Im Gegensatz zum M/M/1-Wartesystem besitzt die Bedienstation c identische parallel arbeitende Bedieneinheiten, die alle

die gleiche Bedienrate $\mu > 0$ haben. Desweiteren ist ein M/M/c-Wartesystem dadurch charakterisiert, daß es

- eine unbeschränkte Systemkapazität,
- eine unbeschränkte Anzahl von potentiellen Kunden,
- die Warteschlangendisziplin *FCFS*

hat.

Zur Analyse eines Warteschlangenmodells vom Typ M/M/c genügt die Kenntnis der Ankunftsrate λ und der Bedienrate μ . Der Zustand eines M/M/c-Wartesystems ist durch die Anzahl der Kunden innerhalb des M/M/c-Wartesystems gegeben, d.h. der Zustandsraum $\mathcal{Z} = \{0, 1, 2, \dots, \infty\}$ ist unendlich. Ein M/M/c-Wartesystem ist somit als ein Spezialfall der Geburts-/Sterbeprozesse (\diamond Abschnitt 4.2.3.3) anzusehen, wobei für die Geburts- und Sterberaten gilt:

$$\begin{aligned} \lambda_i &= \lambda, & \lambda &> 0, \forall i \geq 0, \\ \mu_i &= \min\{i \cdot \mu, c \cdot \mu\} = \begin{cases} i \cdot \mu, & 1 \leq i < c \\ c \cdot \mu, & i \geq c \end{cases} \end{aligned} \quad (4.27)$$

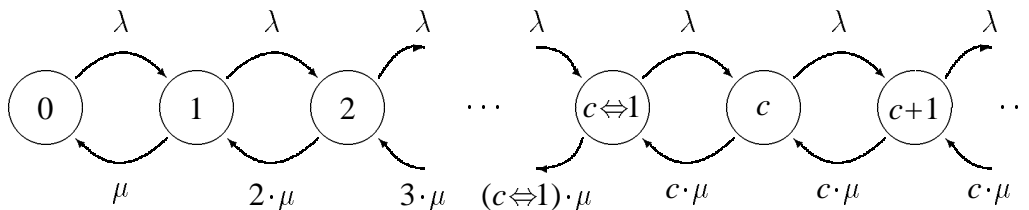


Abbildung 4.9: Zustandstransitionsdiagramm eines M/M/c-Wartesystems

Die Abbildung 4.9 zeigt das Zustandstransitionsdiagramm eines M/M/c-Wartesystems. Damit sich ein M/M/c-Wartesystem im statistischen Gleichgewicht befindet, muß gemäß Gleichung (4.15) und (4.20) für die Auslastung ρ gelten:

$$\rho = \frac{\lambda}{c \cdot \mu} < 1. \quad (4.28)$$

Unter Benutzung von Gleichung (4.27) lassen sich die stationären Zustandswahrscheinlichkeiten eines M/M/c-Wartesystems, also die Wahrscheinlichkeiten p_i , daß sich i Kunden im System befinden, durch Einsetzen in die Gleichung (4.17a) mittels Fallunterscheidung berechnen:

- $1 \leq i < c$:

$$p_i = p_0 \cdot \prod_{k=0}^{i-1} \left(\frac{\lambda}{(k+1) \cdot \mu} \right) = p_0 \cdot \frac{1}{i!} \cdot \left(\frac{\lambda}{\mu} \right)^i \quad (4.29a)$$

- $i \geq c$:

$$p_i = p_0 \cdot \left(\prod_{k=0}^{c-1} \left(\frac{\lambda}{(k+1) \cdot \mu} \right) \right) \cdot \left(\prod_{k=c}^{i-1} \frac{\lambda}{c \cdot \mu} \right) = p_0 \cdot \left(\frac{\lambda}{\mu} \right)^i \cdot \frac{1}{c! \cdot c^{i-c}} \quad (4.29b)$$

Faßt man die Ergebnisse der Gleichungen (4.29a) und (4.29b) zusammen, so lassen sich die stationären Zustandswahrscheinlichkeiten eines M/M/c-Wartesystems durch:

$$p_i = \begin{cases} p_0 \cdot \frac{1}{i!} \cdot \left(\frac{\lambda}{\mu}\right)^i & , 1 \leq i < c \\ p_0 \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{c! \cdot c^{i-c}} & , i \geq c \end{cases} \quad (4.29c)$$

lösen.

Die Initialwahrscheinlichkeit p_0 ist durch die Normalisierungsbedingung für Wahrscheinlichkeiten ($\sum_{i \in \mathbb{Z}} p_i = 1$) und durch Einsetzen der Gleichung (4.29c) bestimmbar.

$$\begin{aligned} 1 &= \sum_{i=0}^{\infty} p_i \stackrel{(4.29c)}{=} \sum_{k=0}^{c-1} p_k + \sum_{n=c}^{\infty} p_k \\ &= \sum_{k=0}^{c-1} p_0 \cdot \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu}\right)^k + \sum_{k=c}^{\infty} p_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{c! \cdot c^{k-c}} \\ &= p_0 \cdot \left(\sum_{k=0}^{c-1} \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu}\right)^k + \sum_{k=c}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{c! \cdot c^{k-c}} \right) \end{aligned} \quad (4.29d)$$

Wird die zweite Summe der Gleichung (4.29d) genauer betrachtet, so gilt:

$$\begin{aligned} \sum_{k=c}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{c! \cdot c^{k-c}} &= \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \sum_{k=c}^{\infty} \left(\frac{\frac{\lambda}{\mu}}{c}\right)^{k-c} \\ &= \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \underbrace{\sum_{k'=0}^{\infty} \left(\frac{\lambda}{c \cdot \mu}\right)^{k'}}_{\text{geometrische Reihe konvergiert, da } \left|\frac{\lambda}{c \cdot \mu}\right| < 1} \\ &= \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \left(\frac{1}{1 \Leftrightarrow \frac{\lambda}{c \cdot \mu}}\right) \\ &= \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \left(\frac{c \cdot \mu}{c \cdot \mu \Leftrightarrow \lambda}\right) . \end{aligned} \quad (4.29e)$$

Somit ist die Initialwahrscheinlichkeit p_0 eines M/M/c-Wartesystems:

$$p_0 = \left[\sum_{k=0}^{c-1} \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu}\right)^k + \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \left(\frac{c \cdot \mu}{c \cdot \mu \Leftrightarrow \lambda}\right) \right]^{-1} . \quad (4.29f)$$

Die stationären Zustandswahrscheinlichkeiten bilden, wie auch bei den M/M/1-Wartesystemen, die Grundlage zur Berechnung der anderen Leistungsgrößen.

Als erstes soll die *mittlere Warteschlangenlänge* $E[n_q]$ ermittelt werden. Mit der Definition des Mittelwertes einer diskreten Zufallsvariablen (\Leftrightarrow Gleichung (4.2)) gilt für die mittlere Warteschlangenlänge eines M/M/c-Wartesystems:

$$E[n_q] = \sum_{i=c}^{\infty} (i \Leftrightarrow c) \cdot p_i = \sum_{i=c}^{\infty} (i \Leftrightarrow c) \cdot \left(p_0 \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{c! \cdot c^{i-c}} \right) . \quad (4.30a)$$

In [Gro74] wird aus der Gleichung (4.30a) hergeleitet, daß sich die mittlere Warteschlangenlänge nach

$$E[n_q] = \left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \lambda \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 \quad (4.30b)$$

berechnen läßt.

Mit dem Gesetz von Little läßt sich aus der mittleren Warteschlangenlänge $E[n_q]$ die *mittlere Wartezeit* $E[q]$ eines Kunden ermitteln:

$$E[q] \stackrel{(4.23b)}{=} \frac{E[n_q]}{\lambda} = \left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 \cdot \quad (4.30c)$$

Die Bedienzeit in einem M/M/c-Wartesystem ist exponentiell verteilt mit Rate μ . Für die *mittlere Bedienzeit* $E[s]$ gilt daher:

$$E[s] = \frac{1}{\mu} \cdot \quad (4.30d)$$

Mit der Summe aus der mittleren Wartezeit $E[q]$ und der mittleren Bedienzeit $E[s]$ ist die *mittlere Verweilzeit* $E[r]$ berechenbar:

$$E[r] = E[q] + E[s] = \left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 + \frac{1}{\mu} \cdot \quad (4.30e)$$

Die *mittlere Kundenanzahl* $E[n]$ kann durch Anwendung des Gesetzes von Little errechnet werden:

$$E[n] \stackrel{(4.23a)}{=} \lambda \cdot E[r] = \left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \lambda \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 + \frac{\lambda}{\mu} \cdot \quad (4.30f)$$

Da die mittlere Kundenanzahl $E[n]$ die Summe aus der mittleren Warteschlangenlänge $E[n_q]$ und der *Anzahl der in Bedienung befindlichen Kunden* ist, kann letztere unter Benutzung der Gleichungen (4.30f) und (4.30b) durch

$$\begin{aligned} E[n_s] &= E[n] \Leftrightarrow E[n_q] \\ &= \left(\left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \lambda \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 + \frac{\lambda}{\mu} \right) \Leftrightarrow \left(\left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \lambda \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 \right) \\ &= \frac{\lambda}{\mu} \end{aligned} \quad (4.30g)$$

berechnet werden.

Eine weitere interessante Größe bei der Analyse von M/M/c-Wartesystemen ist die *Wahrscheinlichkeit*, daß ein Kunde auf seine Bedienung warten muß. Das ist der Fall, wenn sich bereits *mindestens* c Kunden im Wartesystem befinden. Sei n die Zufallsvariable, die die Kundenanzahl eines M/M/c-Wartesystems angibt. Dann gilt:

$$P\{n \geq c\} = \sum_{k=c}^{\infty} p_k$$

$$\begin{aligned}
&\stackrel{(4.29b)}{=} \sum_{k=c}^{\infty} p_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{c! \cdot c^{k-c}} \\
&= p_0 \cdot \sum_{k=c}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{c! \cdot c^{k-c}} \\
&\stackrel{(4.29e)}{=} p_0 \cdot \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \left(\frac{c \cdot \mu}{c \cdot \mu \Leftrightarrow \lambda}\right).
\end{aligned}$$

Die Tabelle A.2 im Anhang gibt einen Überblick über die in diesem Abschnitt hergeleiteten Formeln für die Analyse von M/M/c-Wartesystemen.

In den Abschnitten 4.3.3 und 4.3.4 wurden die Wartesysteme vom Typ M/M/1 und M/M/c vorgestellt. Beiden Wartesystemen ist gemein, daß

- die Zwischenankunftszeiten exponentiell verteilt sind und der Ankunftsprozeß ein Poisson-Prozeß ist,
- die Bedienzeiten exponentiell verteilt sind,
- die Systemkapazität unbeschränkt ist,
- die Anzahl von potentiellen Kunden unbeschränkt ist,
- die Warteschlangendisziplin *FCFS* ist.

Die vorgestellten Wartesysteme unterscheiden sich in der Anzahl der Bedieneinheiten innerhalb der Bedienstation. So hat ein M/M/1-Wartesystem nur eine Bedieneinheit, während ein M/M/c-Wartesystem c identische parallel arbeitende Bedieneinheiten besitzt.

Aufgrund der Warteschlangendisziplin *FCFS* werden *alle* ankommenden Kunden in der Reihenfolge ihrer Ankunft bedient. Oftmals ist es aber so, daß bestimmte Kunden *bevorzugt* bedient werden müssen. Sie werden dann bei ihrer Ankunft einer bestimmten Prioritätsklasse zugeordnet. Die Bedienstrategie ist eine statische Prioritäts-Disziplin. Die Wartesysteme vom Typ M/M/c werden im nächsten Abschnitt um statische Prioritäts-Disziplinen erweitert.

4.3.5 M/M/c-Wartesysteme mit statischer Prioritäts-Disziplin

Im Abschnitt 4.1 wurden verschiedene Arten von Bedienstrategien beschrieben. Die Bedienstrategie nach statischen Prioritäten soll in diesem Abschnitt für M/M/c-Wartesysteme beschrieben werden. Dabei wird zwischen *unterbrechenden* (preemptive) und *nichtunterbrechenden* (nonpreemptive) Prioritäts-Disziplinen unterschieden [Bol89]:

- **M/M/c-FCFS PR** (preemptive)
Es handelt sich um eine *unterbrechende* Strategie. Ein gerade in Bedienung befindlicher Kunde wird bei Ankunft eines Kunden höherer Priorität unterbrochen. Wegen der exponentiellen Bedienzeitverteilung ist es *nicht* von Bedeutung, ob die unterbrochene Bedienung des Kunden niedrigerer Priorität nach Abarbeitung der Kunden höherer Priorität neu

gestartet oder an der Unterbrechungsstelle fortgeführt wird. Die Bedienstrategie innerhalb einer Prioritätsklasse ist *FCFS*.

- **M/M/c-FCFS NPR** (nonpreemptive)

Diese Strategie ist *nichtunterbrechend*. Erst nach Abarbeitung des gerade in Bedienung befindlichen Kunden werden Kunden einer höheren Prioritätsklasse bedient. Die Abarbeitungsstrategie innerhalb einer Klasse ist *FCFS*.

Die Menge von Prioritätsklassen $\mathcal{P} = \{1, 2, \dots, k\}$ ist linear geordnet, d.h. bei k Prioritätsklassen hat die Klasse 1 die höchste Priorität und die Klasse k die niedrigste Priorität.

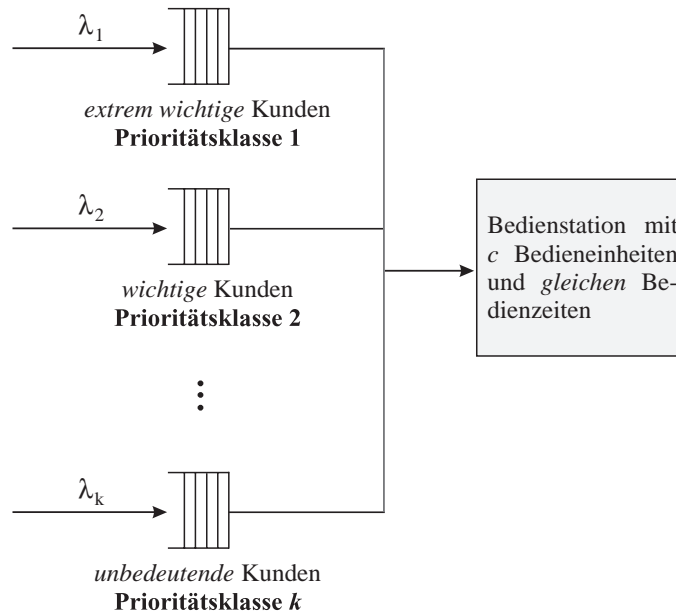


Abbildung 4.10: M/M/c-Wartesystem mit Prioritäten

4.3.5.1 Das M/M/c-FCFS PR-Wartesystem

In [BB83] wird die mittlere Verweilzeit eines M/M/c-FCFS PR-Wartesystems untersucht. Dabei wird von einem M/M/c-Wartesystem mit k Prioritätsklassen ausgegangen, welches die folgenden Annahmen erfüllt:

- Kunden der Prioritätsklasse i haben bei der Bedienung Vorrang vor Kunden der Klasse j , falls $1 \leq i < j \leq k$.
- Für Kunden innerhalb der gleichen Prioritätsklasse gilt die Warteschlangendisziplin *FCFS*.
- Der Ankunftsprozeß für Kunden der Prioritätsklasse i ist ein Poisson-Prozeß mit Rate λ_i . Die Bedienzeiten der einzelnen Prioritätsklassen sind exponentiell verteilt mit Rate μ_i und es gilt $\mu = \mu_i, \forall i \in \mathcal{P}$.

Zur weiteren Untersuchung der mittleren Verweilzeit eines M/M/c-FCFS PR-Wartesystems werden einige Schreibweisen eingeführt.

$\lambda_{(p)}$ Aggregierte Ankunftsrate der Prioritätsklasse p :

$$\lambda_{(p)} = \sum_{i=1}^p \lambda_i, \quad p \in \mathcal{P}$$

$E[r]_i$ Mittlere Verweilzeit für Kunden der Prioritätsklasse i

$E[r]_{(p)}$ Aggregierte mittlere Verweilzeit der Prioritätsklasse p . Diese wird berechnet durch:

$$E[r]_{(p)} = \frac{\sum_{i=1}^p \lambda_i \cdot E[r]_i}{\lambda_{(p)}}, \quad p \in \mathcal{P}. \quad (4.31)$$

Damit sich das M/M/c-FCFS PR-Wartesystem im statistischen Gleichgewicht befindet, muß für die Auslastung

$$\rho^{(k)} = \frac{\sum_{i=1}^k \lambda_i}{c \cdot \mu} = \frac{\lambda^{(k)}}{c \cdot \mu} < 1$$

gelten.

Um nun die mittleren Verweilzeiten für die einzelnen Prioritätsklassen $E[r]_i$ und damit auch die aggregierte mittlere Verweilzeit $E[r]_{(p)}$ berechnen zu können, wird in [Bar80] folgender Ansatz für $p \geq 2$ gewählt:

$$\begin{aligned} E[r]_{(p)} &\stackrel{(4.31)}{=} \frac{\sum_{i=1}^p \lambda_i \cdot E[r]_i}{\lambda_{(p)}} \\ &= \frac{\sum_{i=1}^{p-1} \lambda_{(p-1)} \cdot \lambda_i \cdot E[r]_i}{\lambda_{(p-1)} \cdot \lambda_{(p)}} + \frac{\lambda_p \cdot E[r]_p}{\lambda_{(p)}} \\ &= \frac{\lambda_{(p-1)}}{\lambda_{(p)}} \cdot \left(\frac{\sum_{i=1}^{p-1} \lambda_i \cdot E[r]_i}{\lambda_{(p-1)}} \right) + \frac{\lambda_p \cdot E[r]_p}{\lambda_{(p)}} \\ &\stackrel{(4.31)}{=} \frac{\lambda_{(p-1)} \cdot E[r]_{(p-1)}}{\lambda_{(p)}} + \frac{\lambda_p \cdot E[r]_p}{\lambda_{(p)}} \quad \Leftrightarrow \\ E[r]_p &= \frac{\lambda_{(p)} \cdot E[r]_{(p)} \Leftrightarrow \lambda_{(p-1)} \cdot E[r]_{(p-1)}}{\lambda_p} \end{aligned} \quad (4.32)$$

Zur Berechnung der Gleichung (4.32) ist die Bestimmung der Werte für $E[r]_{(p)}$ und $E[r]_{(p-1)}$ notwendig. Dabei sind zwei Feststellungen hilfreich:

1. In [Kle76] wird festgestellt, daß die Verweilzeit der höchsten p Prioritätsklassen von der Anwesenheit von Kunden niedrigerer Prioritätsklassen nicht beeinflußt wird.
2. In [Jai68] wird festgestellt, daß sich die aggregierte mittlere Verweilzeit in einem beliebigen Wartesystem, in dem alle Kundenklassen exponentiell verteilte Bedienzeiten mit gleicher

Rate μ haben, bei einem Wechsel der Warteschlangendisziplin von *FCFS* nach *FCFS PR* nicht ändert.

Aus diesen beiden Feststellungen läßt sich folgern, daß die aggregierte mittlere Verweilzeit $E[r]_{(p)}$ der ersten p Prioritätsklassen gleich der mittleren Verweilzeit von Kunden in einem M/M/c-Wartesystem mit Ankunftsrate $\lambda_{(p)}$ und Bedienrate μ für $p \in \mathcal{P}$ ist. Somit bleibt die mittlere Verweilzeit für die ersten p Prioritätsklassen bei einem Wechsel der Bedienstrategie nach *FCFS* und der Elimination von Kunden der Prioritätsklassen $> p$ gleich.

Mit diesen Feststellungen lassen sich die Werte für $E[r]_{(p)}$ und $E[r]_{(p-1)}$ durch die im Abschnitt 4.3.4 für M/M/c-Wartesysteme hergeleiteten Formeln ermitteln.

Setzt man $\rho = \frac{\lambda}{c \cdot \mu}$ in Gleichung (4.29f) ein, so gilt für die Initialwahrscheinlichkeit eines M/M/c-Wartesystems:

$$p_0 = \left[\sum_{k=0}^{c-1} \frac{(c \cdot \rho)^k}{k!} + \left(\frac{(c \cdot \rho)^c}{c!} \right) \cdot \left(\frac{1}{1 \Leftrightarrow \rho} \right) \right]^{-1} \quad (4.33a)$$

Wird die gleiche Ersetzung in Gleichung (4.30e) gemacht, dann läßt sich die mittlere Verweilzeit eines M/M/c-Wartesystems auch durch

$$E[r] = \frac{\rho \cdot (c \cdot \rho)^c}{\lambda \cdot c! \cdot (1 \Leftrightarrow \rho)^2} \cdot p_0 + \frac{1}{\mu} \quad (4.33b)$$

berechnen.

Ersetzt man in Gleichung (4.33b)

$$\begin{aligned} \lambda &= \lambda_{(p)} \\ \rho &= \rho_{(p)} = \frac{\lambda_{(p)}}{c \cdot \mu} \\ p_0 &= p_0(\rho_{(p)}) \end{aligned}$$

dann erhält man für $E[r]_{(p)}$:

$$E[r]_{(p)} = \frac{\rho_{(p)} \cdot (c \cdot \rho_{(p)})^c}{\lambda_{(p)} \cdot c! \cdot (1 \Leftrightarrow \rho_{(p)})^2} \cdot p_0(\rho_{(p)}) + \frac{1}{\mu}, \quad p \geq 2 \quad (4.34a)$$

und für $E[r]_{(p-1)}$:

$$E[r]_{(p-1)} = \frac{\rho_{(p-1)} \cdot (c \cdot \rho_{(p-1)})^c}{\lambda_{(p-1)} \cdot c! \cdot (1 \Leftrightarrow \rho_{(p-1)})^2} \cdot p_0(\rho_{(p-1)}) + \frac{1}{\mu}, \quad p \geq 2. \quad (4.34b)$$

Werden die Gleichungen (4.34a) und (4.34b) in die Gleichung (4.32) eingesetzt, so gilt für $E[r]_p$:

$$\begin{aligned} E[r]_p &= \frac{\lambda_{(p)} \cdot \left(\frac{\rho_{(p)} \cdot (c \cdot \rho_{(p)})^c}{\lambda_{(p)} \cdot c! \cdot (1 \Leftrightarrow \rho_{(p)})^2} \cdot p_0(\rho_{(p)}) + \frac{1}{\mu} \right) \Leftrightarrow \lambda_{(p-1)} \cdot \left(\frac{\rho_{(p-1)} \cdot (c \cdot \rho_{(p-1)})^c}{\lambda_{(p-1)} \cdot c! \cdot (1 \Leftrightarrow \rho_{(p-1)})^2} \cdot p_0(\rho_{(p-1)}) + \frac{1}{\mu} \right)}{\lambda_p} \\ &= \frac{1}{\mu} + \frac{\rho_{(p)} \cdot (c \cdot \rho_{(p)})^c}{\lambda_p \cdot c! \cdot (1 \Leftrightarrow \rho_{(p)})^2} \cdot p_0(\rho_{(p)}) \Leftrightarrow \\ &\quad \frac{\rho_{(p-1)} \cdot (c \cdot \rho_{(p-1)})^c}{\lambda_p \cdot c! \cdot (1 \Leftrightarrow \rho_{(p-1)})^2} \cdot p_0(\rho_{(p-1)}) , \quad p \geq 2. \end{aligned} \quad (4.35a)$$

Für $p = 1$ ergibt sich $E[r]_{(1)}$ direkt aus Gleichung (4.33b), wenn man $\lambda = \lambda_{(1)}, \rho = \rho_{(1)}$ und $p_0 = p_0(\rho_{(1)})$ setzt:

$$E[r]_{(1)} = E[r]_1 = \frac{\rho_{(1)} \cdot (c \cdot \rho_{(1)})^c}{\lambda_{(1)} \cdot c! \cdot (1 \Leftrightarrow \rho_{(1)})^2} \cdot p_0(\rho_{(1)}) + \frac{1}{\mu} . \quad (4.35b)$$

Durch die Gleichungen (4.35) ist man in der Lage, die mittleren Verweilzeiten für alle Prioritätsklassen und die aggregierte mittlere Verweilzeit des M/M/c-FCFS PR-Wartesystems nach Gleichung (4.31) für $p = k$ zu berechnen. Diese ist nach den Feststellungen auf Seite 45 gleich der mittleren Verweilzeit in einem M/M/c-Wartesystem.

Mit den oben hergeleiteten Formeln lassen sich auch andere Leistungsgrößen durch Benutzen des Gesetz von Little (\diamond Abschnitt 4.3.2) und den Beziehungen $E[n] = E[n_q] + E[n_s]$ bzw. $E[r] = E[q] + E[s]$ für die einzelnen Prioritätsklassen und als aggregierte Werte berechnen. So ist beispielsweise die *mittlere Kundenanzahl* der Prioritätsklasse p nach Gleichung (4.23a) durch

$$E[n]_p = \lambda_p \cdot E[r]_p \quad (4.36a)$$

berechenbar und die *aggregierte mittlere Kundenanzahl* $E[n]_{(p)}$ der höchsten p Prioritätsklassen läßt sich mit

$$E[n]_{(p)} = \sum_{i=1}^p E[n]_i \quad (4.36b)$$

berechnen. Für $p = k$ erhält man die aggregierte mittlere Kundenanzahl des M/M/c-FCFS PR-Wartesystems aus Gleichung (4.36b).

4.3.5.2 Das M/M/c-FCFS NPR-Wartesystem

Bei der Analyse eines M/M/c-FCFS NPR-Wartesystem wird von einem M/M/c-Wartesystem mit k Prioritätsklassen ausgegangen, welches die gleichen Annahmen erfüllt wie im Abschnitt 4.3.5.1. Im Gegensatz zu den M/M/c-FCFS PR-Wartesystemen wird ein in der Bedieneinheit $i \leq c$ befindlicher Kunde, egal welcher Prioritätsklasse $p \in \mathcal{P} = \{1, 2, \dots, k\}$ er angehört, erst komplett abgearbeitet bevor ein anderer Kunde die Bedieneinheit i benutzen kann.

Sei

$\lambda_{(p)}$ die aggregierte Ankunftsrate der Prioritätsklasse p :

$$\lambda_{(p)} = \sum_{i=1}^p \lambda_i, \quad p \in \mathcal{P},$$

ρ_p die Auslastung der Prioritätsklasse p :

$$\rho_p = \frac{\lambda_p}{c \cdot \mu}, \quad p \in \mathcal{P},$$

$\rho^{(k)}$ Die Auslastung des M/M/c-FCFS NPR-Wartesystems:

$$\rho^{(k)} = \frac{\sum_{i=1}^k \lambda_i}{c \cdot \mu} = \frac{\lambda^{(k)}}{c \cdot \mu}.$$

Damit sich das M/M/c-FCFS NPR-Wartesystem im statistischen Gleichgewicht befindet, muß für die Auslastung

$$\rho^{(k)} < 1$$

gelten.

Wenn ein Kunde der Prioritätsklasse p zum Zeitpunkt t_0 am Wartesystem ankommt, und zum Zeitpunkt t_1 von einer freien Bedienstation zur Bedienung übernommen wird, ist die Wartezeit für diesen Kunden $q = t_1 \Leftrightarrow t_0$. Zum Zeitpunkt t_0 sind

- n_1 Kunden der Prioritätsklasse 1
- n_2 Kunden der Prioritätsklasse 2
- \vdots
- n_p Kunden der Prioritätsklasse p

in der Warteschlange, die vor dem betrachteten Kunden bedient werden.

Sei s_0 die Zeit die vergeht, bis die nächste Bedieneinheit frei wird und s_j die Zeit die benötigt wird um n_j Kunden zu bedienen ($j \geq 1$). In der Zeit, in der der betrachtete Kunde in der Warteschlange auf seine Bedienung wartet, erreichen n'_j Kunden der Prioritätsklassen $j < p$ das Wartesystem und werden *vor* dem betrachteten Kunden bedient. Die Zeit die dafür benötigt wird sei s'_j . Somit läßt sich die Wartezeit eines Kunden der Prioritätsklasse p durch

$$q = \sum_{j=1}^{p-1} s'_j + \sum_{j=1}^p s_j + s_0, \quad p \in \mathcal{P},$$

berechnen.

Stellt man diesen Ausdruck in Erwartungswerten dar, dann gilt:

$$E[q]_p = \sum_{j=1}^{p-1} E[s'_j] + \sum_{j=1}^p E[s_j] + E[s_0], \quad p \in \mathcal{P}. \quad (4.37)$$

In [Gro74] wird hergeleitet, daß für die mittlere Wartezeit eines Kunden der Prioritätsklasse p nach Gleichung (4.37) gilt:

$$E[q]_p = \frac{E[s_0]}{(1 \Leftrightarrow \rho_{(p-1)}) \cdot (1 \Leftrightarrow \rho_{(p)})}, \quad p \in \mathcal{P},$$

wobei

$$E[s_0] = \frac{(c \cdot \rho^{(k)})^c}{c! \cdot (1 \Leftrightarrow \rho^{(k)}) \cdot (c \cdot \mu)} \cdot \left[\sum_{i=0}^{c-1} \frac{(c \cdot \rho^{(k)})^i}{i!} + \frac{(c \cdot \rho^{(k)})^c}{c! \cdot (1 \Leftrightarrow \rho^{(k)})} \right]^{-1}$$

ist.

Insgesamt ist die mittlere Wartezeit für Kunden der Prioritätsklassen $p \leq k$:

$$E[q]_p = \frac{\left[c! \cdot (1 \Leftrightarrow \rho_{(k)}) \cdot (c \cdot \mu) \cdot \sum_{i=0}^{c-1} \frac{(c \cdot \rho_{(k)})^{i-c}}{i!} + c \cdot \mu \right]^{-1}}{(1 \Leftrightarrow \rho_{(p-1)}) \cdot (1 \Leftrightarrow \rho_{(p)})}, \quad p \in \mathcal{P}. \quad (4.38a)$$

Aus den mittleren Wartezeiten für die einzelnen Prioritätsklassen $p \leq k$ läßt sich die *aggregierte mittlere Wartezeit* für die p höchsten Prioritätsklassen durch

$$E[q]_{(p)} = \frac{\sum_{i=1}^k \lambda_i \cdot E[q]_i}{\lambda_{(k)}}, \quad p \in \mathcal{P}, \quad (4.38b)$$

berechnen. Für $p = k$ erhält man die aggregierte mittlere Wartezeit des M/M/c-FCFS NPR-Wartesystems, die gleich der mittleren Wartezeit eines M/M/c-Wartesystems ist.

Wie im Falle des M/M/c-FCFS PR-Wartesystems lassen sich andere Leistungsgrößen durch Benutzen des Gesetz von Little (\Leftrightarrow Abschnitt 4.3.2) und den Beziehungen $E[n] = E[n_q] + E[n_s]$ bzw. $E[r] = E[q] + E[s]$ für die einzelnen Prioritätsklassen und als aggregierte Werte berechnen. So ist beispielsweise die mittlere Warteschlangenlänge der Prioritätsklasse p nach Gleichung (4.23b) durch

$$E[n_q]_p = \lambda_p \cdot E[q]_p \quad (4.39a)$$

berechenbar und die *aggregierte mittlere Warteschlangenlänge* $E[n_q]_{(p)}$ der höchsten p Prioritätsklassen läßt sich mit

$$E[n_q]_{(p)} = \sum_{i=1}^p E[n_q]_i \quad (4.39b)$$

berechnen. Für $p = k$ erhält man die aggregierte mittlere Warteschlangenlänge des M/M/c-FCFS NPR-Wartesystems aus Gleichung (4.39b).

Die hergeleiteten Formeln für M/M/c-Wartesysteme mit statischen Prioritäts-Disziplinen lassen sich auch für M/M/1-Wartesysteme mit statischen Prioritäts-Disziplinen anwenden, indem $c = 1$ gesetzt wird.

Im Abschnitt 4.3 wurden einige elementare Markov'sche Wartesysteme und Grundlagen zur Berechnung diverser Leistungsgrößen vorgestellt. Es gibt noch eine Vielzahl anderer Warteschlangensysteme. Diese alle hier vorzustellen würde jedoch den Rahmen dieser Arbeit sprengen. Es wurden nur solche Wartesysteme vorgestellt, die in dieser Arbeit von Interesse sind.

Bei der Leistungsbewertung mittels Warteschlangenmodellen gilt das Interesse oftmals nicht nur einzelnen Wartesystemen. Bei den meisten praktischen Anwendungen werden Systeme betrachtet, in denen an mehreren Stellen ein Bedienvorgang erfolgt. *Netze* von elementaren Wartesystemen liefern eine mathematische Grundlage für die Modellierung und Analyse solcher Systeme.

4.4 Warteschlangennetze

Bisher wurden nur einzelne Warteschlangensysteme betrachtet. In diesem Abschnitt werden Warteschlangenmodelle beschrieben, die sich aus mehreren elementaren Wartesystemen zusammensetzen. Solche Warteschlangenmodelle werden *Warteschlangennetze* genannt, wenn mindestens zwei elementare Wartesysteme miteinander verbunden sind. Die elementaren Wartesysteme innerhalb eines Warteschlangennetzes werden auch *Netzknotten* oder einfach nur *Knotten* genannt. Ein Netzknotten repräsentiert eine aktive Komponente in einem realen bzw. zu modellierenden System. Übergänge von Kunden sind prinzipiell zwischen allen Knotten des Netzes möglich. Insbesondere auch die direkte Rückführung eines Kunden zum gleichen Knotten.

Ein Warteschlangennetz ist *offen*, wenn Kunden von außerhalb des Netzes durch eine *Quelle* ankommen und dieses auch durch eine *Senke* wieder verlassen können. Im Gegensatz dazu wird ein Warteschlangennetz *geschlossen* genannt, wenn keine externen Ankünfte und Abgänge von Kunden möglich sind. Bei geschlossenen Netzen ist die Kundenanzahl konstant. Eine zusätzliche Verfeinerung der Warteschlangennetze ist durch die Einführung verschiedener Kundenklassen möglich. Die Kundenklassen unterscheiden sich durch unterschiedliche Zwischenankunfts- und Bedienzeiten. Enthält ein Warteschlangennetz sowohl offene als auch geschlossene Kundenklassen, so wird es *gemischt* genannt.

Die in diesem Abschnitt betrachteten Warteschlangennetze erfüllen *alle* die folgenden Annahmen:

- Das Warteschlangennetz besteht aus einer Menge $\mathcal{N} = \{1, 2, \dots, N\}$ von Knotten. Die Kardinalität der Knottenmenge \mathcal{N} ist $\text{card}(\mathcal{N}) \geq 2$.
- Jeder Knotten $i, i \in \mathcal{N}$, hat einen exponentiell verteilten Bedienprozeß mit Rate μ_i .
- Der externe Ankunftsprozeß des Knotten $i, i \in \mathcal{N}$, ist ein Poisson-Prozeß mit Ankunftsrate λ_{0i} .
- Die Warteschlangendisziplin der einzelnen Knotten ist *FCFS*.
- Befindet sich ein Kunde im Knotten i , dann geht er mit der Wahrscheinlichkeit p_{ij} zum Knotten j über, $i, j \in \mathcal{N}$. Ist der Bedienvorgang eines Kunden im Knotten i beendet, so verläßt er das Warteschlangennetz mit der Wahrscheinlichkeit $1 \Leftrightarrow \sum_{j=1}^N p_{ij}$.

Ein Warteschlangennetz, welches nur Knotten des gerade beschriebenen Typs besitzt, besteht aus N M/M/c-Wartesystemen. In solchen Systemen wird das Abgangsverhalten der Kunden an den einzelnen Knotten nach dem Theorem von Burke [Bur59] durch einen Poisson-Prozeß beschrieben, wenn sich das Wartesystem im stationären Zustand (statistischen Gleichgewicht) befindet [Bur68]. Die Rate des Abgangsprozesses ist dann gleich der Ankunftsrate λ_i .

Damit sich alle Knotten im statistischen Gleichgewicht befinden, muß gelten:

$$\rho_i = \frac{\lambda_i}{c_i \cdot \mu_i} < 1, \quad c_i \geq 1, \forall i \in \mathcal{N}. \quad (4.40)$$

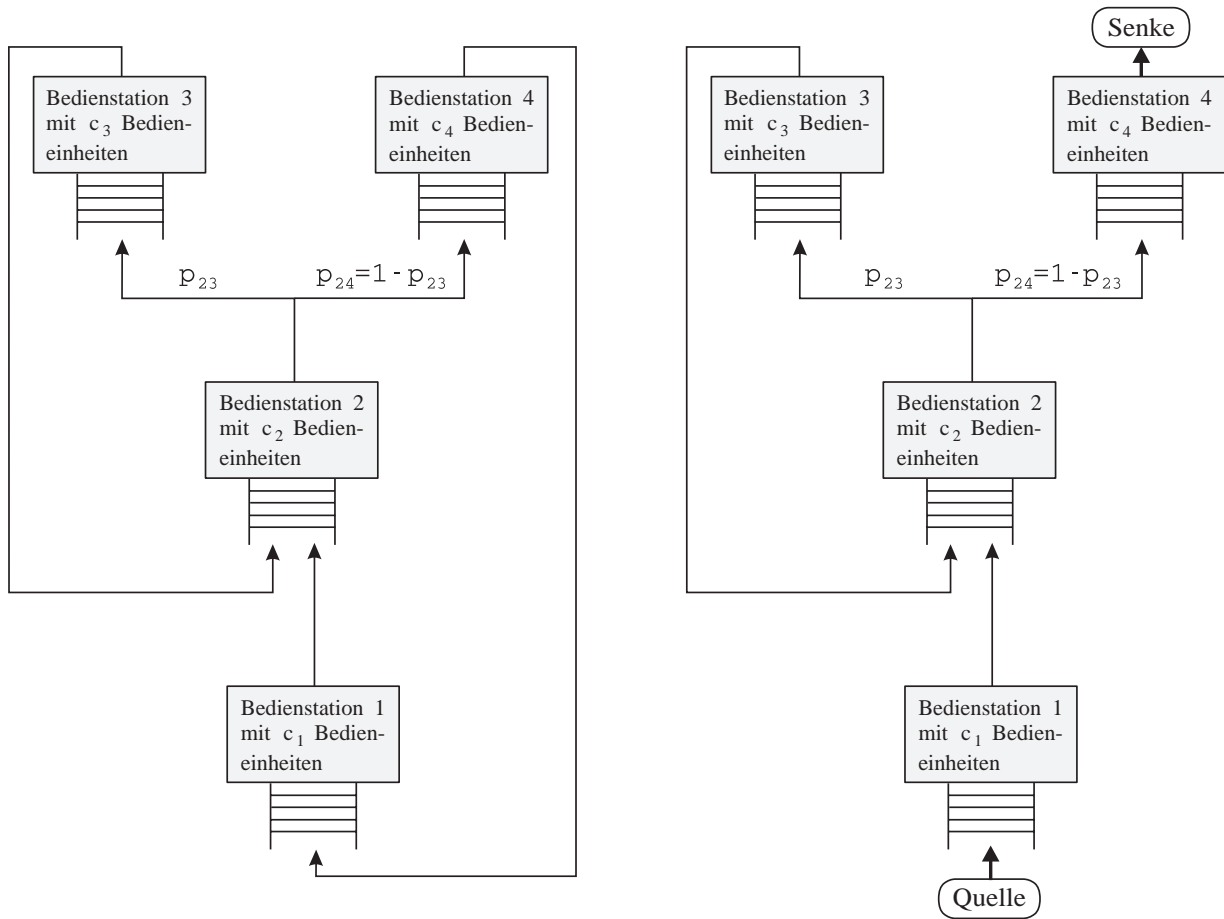


Abbildung 4.11: Geschlossenes und offenes Warternetz

4.4.1 Formale Beschreibung von Warteschlangennetzen

Bei der formalen Beschreibung von Warteschlangennetzen werden die folgenden Bezeichnungen verwendet.

N Anzahl der Netzknoten, wobei $N = \text{card}(\mathcal{N})$.

k_i Anzahl von Kunden im Knoten i , $i \in \mathcal{N}$.

$\vec{k} = (k_1, k_2, \dots, k_N)$ Zustand des Warteschlangennetzes. Das Verhalten des Warteschlangennetzes läßt sich durch eine Markov-Kette beschreiben, in der die Vektoren \vec{k} den Zustandsraum \mathcal{Z} aufspannen.

$p(\vec{k})$ Zustandswahrscheinlichkeit des Warteschlangennetzes.

c_i Anzahl der identischen parallelen Bedieneinheiten in der Bedienstation des Knoten i , $i \in \mathcal{N}$.

μ_i Mittlere Bedienrate des exponentiell verteilten Bedienprozesses im Knoten i , $i \in \mathcal{N}$.

$\frac{1}{\mu_i}$ Mittlere Bedienzeit des Bedienprozesses im Knoten i , $i \in \mathcal{N}$.

p_{ij} Wahrscheinlichkeit, daß ein im Knoten i fertig bedienter Kunde zum Knoten j wechselt (Verzweigungswahrscheinlichkeit), $i, j \in \mathcal{N}$. Für die Verzweigungswahrscheinlichkeiten

gilt:

$$\sum_{j=1}^N p_{ij} = 1, \quad \forall i \in \mathcal{N}. \quad (4.41a)$$

p_{i0} Wahrscheinlichkeit, daß ein Kunde nach der Bedienung im Knoten i das Warteschlangennetz verläßt. Es gilt:

$$p_{i0} = 1 \Leftrightarrow \sum_{j=1}^N p_{ij}, \quad \forall i \in \mathcal{N}.$$

λ_{0i} Mittlere externe Ankunftsrate des Poisson'schen Ankunftsprozesses am Knoten i , $i \in \mathcal{N}$.

λ_i Mittlere Ankunftsrate des Ankunftsprozesses am Knoten i , $i \in \mathcal{N}$.

Zur Berechnung der λ_i für alle Knoten $i \in \mathcal{N}$ müssen die externen Ankünfte und die Ankünfte von allen internen Knoten summiert werden (Φ Abschnitt 4.2.3.5). Da sich das Netz im statistischen Gleichgewicht befindet, ist die mittlere Ankunftsrate eines Knotens gleich den mittleren, mit den entsprechenden Verzweigungswahrscheinlichkeiten gewichteten, Abgangsraten aller Knoten:

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^N p_{ji} \cdot \lambda_j, \quad \forall i \in \mathcal{N}. \quad (4.41b)$$

Durch Lösen dieses Gleichungssystems können die mittleren Ankunftsraten λ_i für *alle* Knoten des Netzes berechnet werden. Die Lösung des Gleichungssystems ist eindeutig, wenn für einige Knoten die Abgangswahrscheinlichkeit $p_{i0} \neq 0$ ist.

Λ Durchsatz eines offenen Netzes:

$$\Lambda = \sum_{i=1}^N \lambda_{0i}.$$

e_i Mittlere Anzahl von Besuchen eines Kunden beim Knoten i , $i \in \mathcal{N}$ (Besuchshäufigkeit):

$$e_i = \frac{\lambda_i}{\Lambda}. \quad (4.41c)$$

x_i Relative Auslastung des Knoten i , $i \in \mathcal{N}$:

$$x_i = \frac{e_i}{\mu_i}.$$

4.4.2 Produktform-Warteschlangennetze

Produktform-Warteschlangennetze haben den großen Vorteil, daß sie eine exakte Bestimmung der Leistungsgrößen und damit eine aussagekräftige Analyse eines Warteschlangennetzes zulassen. Diese Warteschlangennetze sind dadurch charakterisiert, daß sich die Lösungen für die Zustandswahrscheinlichkeiten \vec{k} des Warteschlangennetzes im statistischen Gleichgewicht multiplikativ aus den stationären Zustandswahrscheinlichkeiten der einzelnen Netzknoten $p_i(k_i)$, $i \in \mathcal{N}$, zusammensetzen (Produktformlösung).

Eine notwendige und hinreichende Bedingung für die Existenz einer *Produktformlösung* ist die *Local-Balance-Eigenschaft* [Spa92]. Die Local-Balance-Eigenschaft besagt:

Ein Warteschlangennetz befindet sich genau dann im lokalen Gleichgewicht, wenn an allen Knoten i , $i \in \mathcal{N}$, die Abgangsrate gleich der Ankunftsrate für einen beliebigen Zustand ist.

Besitzt ein Warteschlangennetz die Local-Balance-Eigenschaft, dann existiert für dieses Warteschlangennetz die Produktformlösung:

$$p(\vec{k}) = p(k_1, k_2, \dots, k_N) = \frac{1}{G} \cdot \prod_{i=1}^N p_i(k_i). \quad (4.42)$$

Die Lösungen der Zustandswahrscheinlichkeiten des Netzes setzen sich aus den Produkten für die Zustandswahrscheinlichkeiten der einzelnen Netzknoten zusammen. Die Normalisierungskonstante G ist so gewählt, daß sich die Wahrscheinlichkeiten aller Zustände des Warteschlangennetzes zu Eins summieren müssen.

Die Produktformlösung aus Gleichung (4.42) besagt, daß sich in Warteschlangennetzen, die die Local-Balance-Eigenschaft besitzen, die einzelnen Netzknoten so verhalten, als wären sie elementare Wartesysteme. Jeder einzelne Knoten kann isoliert vom Rest des Warteschlangennetzes untersucht werden. Daher werden Produktform-Warteschlangennetze oft auch als *separable Warteschlangennetze* bezeichnet.

In [Cha72] wird gezeigt, daß die unten genannten Typen von elementaren Wartesystemen stets die Local-Balance-Eigenschaft besitzen.

- M/M/c
- M/G/1-PS(RR)
- M/G/ ∞ -IS
- M/G/1-LCFS PR

4.4.2.1 Jackson-Netze

J. R. Jackson hat in seinen Arbeiten [Jac57] und [Jac63] gezeigt, daß offene Warteschlangennetze, deren Knoten alle vom Typ M/M/c sind und sich im statistischen Gleichgewicht befinden, eine Produktformlösung haben.

Satz 4.11: (*Jackson-Theorem für offene Warteschlangennetze [Jac57]*)

Definiere $p_i(k_i)$, $i \in \mathcal{N}$, $k_i \in \mathcal{Z}$, durch die folgenden Gleichungen:

$$p_i(k_i) = \begin{cases} p_i(0) \cdot \left(\frac{\lambda_i}{\mu_i}\right) \cdot \frac{1}{k!} & , 1 \leq k < c_i \\ p_i(0) \cdot \left(\frac{\lambda_i}{\mu_i}\right) \cdot \frac{1}{c_i! \cdot c_i^{k-c_i}} & , k \geq c_i \end{cases}, \quad \forall i \in \mathcal{N}.$$

Die stationären Zustandswahrscheinlichkeiten eines offenen Warteschlangennetzes, das aus M/M/c-Wartesystemen besteht, ist durch das Produkt

$$p(\vec{k}) = p(k_1, k_2, \dots, k_n) = \prod_{i=1}^N p_i(k_i)$$

gegeben. Dabei wird vorausgesetzt, daß $\lambda_i < c_i \cdot \mu_i, \forall i \in \mathcal{N}$, ist.

Das Warteschlangennetz läßt sich also durch die Einzelanalyse der einzelnen Netzknoten analysieren. Die Analyse vollzieht sich in drei Schritten:

1. Für alle Knoten $i, i \in \mathcal{N}$ werden mit dem Gleichungssystem (4.41b) die Ankunftsraten λ_i ermittelt.
2. Jeder Knoten wird als elementares M/M/c-Wartesystem analysiert. Dabei ist die Stabilitätsbedingung (\diamond Gleichung (4.40)) zu prüfen. Danach können die Zustandswahrscheinlichkeiten und Leistungsgrößen gemäß Abschnitt 4.3.4, bzw. bei M/M/1-Wartesystemen gemäß Abschnitt 4.3.3 berechnet werden.
3. Analyse des gesamten Warteschlangennetzes unter Zuhilfenahme der in Schritt 2 ermittelten Ergebnisse.

4.4.2.2 Gordon/Newell-Netze

Ein Spezialfall der Jackson-Wartennetze sind solche Netze, in denen *keine* externen Ankünfte und Abgänge zugelassen sind. Solche geschlossenen Warteschlangennetze wurden von Gordon und Newell untersucht [GN67]. Sie stellten für diese Warteschlangennetze eine Produktformlösung auf und bewiesen deren Existenz. Aus der Geschlossenheit des Netzes ergibt sich, daß sich immer eine konstante Anzahl $K = \sum_{i=1}^N k_i$ Kunden im Warteschlangennetz befindet. Der Zustandsraum \mathcal{Z} ist somit endlich und ergibt sich aus den verschiedenen Möglichkeiten K Kunden auf N Knoten zu verteilen:

$$\text{card}(\mathcal{Z}) = \binom{N + K}{N}.$$

Die Produktformlösung für solche Gordon/Newell-Warteschlangennetze lautet:

$$p(\vec{k}) = \frac{1}{G(K)} \cdot \prod_{i=1}^N \frac{x_i^{k_i}}{\beta_i(k_i)}, \quad (4.43a)$$

mit

$$\beta_i(k_i) = \begin{cases} k_i! & , 1 \leq k_i < c_i \\ c_i! \cdot c_i^{k_i - c_i} & , k_i \geq c_i \end{cases}, \quad \forall i \in \mathcal{N} \quad (4.43b)$$

und der Normalisierungskonstanten $G(K)$, die sich aus der Bedingung ergibt, daß die Summe aller Zustandswahrscheinlichkeiten Eins sein muß:

$$G(K) = \sum_{\vec{k}} \prod_{i=1}^N \frac{x_i^{k_i}}{\beta_i(k_i)}. \quad (4.43c)$$

In der letzten Gleichung wird über alle Zustände $\vec{k} = (k_1, k_2, \dots, k_N)$ summiert, für die

$$\sum_{i=1}^N k_i = K$$

gilt.

Die Analyse eines Gordon/Newell-Netzes kann durch die folgenden sechs Schritte beschrieben werden:

1. Für alle Knoten i , $i \in \mathcal{N}$, des geschlossenen Netzes werden die Besuchshäufigkeiten e_i nach Gleichung (4.41c) berechnet.
2. Die Funktion $\beta_i(k_i)$ wird für alle Knoten i , $i \in \mathcal{N}$ gemäß Gleichung (4.43b) berechnet.
3. Berechnung der Normalisierungskonstanten G nach Gleichung (4.43c).
4. Berechnung der Zustandswahrscheinlichkeiten des gesamten Wartennetzes nach Gleichung (4.43a).
5. Bestimmung der *Randwahrscheinlichkeiten* aus den Zustandswahrscheinlichkeiten des Warteschlangennetzes. Die Randwahrscheinlichkeiten $p_i(k)$ eines geschlossenen Warteschlangennetzes sind die Wahrscheinlichkeiten, daß sich im Knoten i , $i \in \mathcal{N}$, genau $k_i = k$ Kunden befindet. Diese werden durch:

$$p_i(k) = \sum_{\left(\sum_{j=1}^N k_j = K\right) \wedge k_i = k} p(\vec{k})$$

berechnet. Danach können die Leistungsgrößen der einzelnen Netzknoten gemäß Abschnitt 4.3.4, bzw. bei M/M/1-Wartesystemen gemäß Abschnitt 4.3.3 aus den Randwahrscheinlichkeiten berechnet werden.

6. Analyse des gesamten Warteschlangennetzes unter Zuhilfenahme der in Schritt 5 ermittelten Ergebnisse.

4.4.2.3 BCMP-Netze

In [BCMP75] werden die Ergebnisse von Jackson bzw. Gordon/Newell auf Warteschlangennetze mit mehreren Kundenklassen, verschiedenen Warteschlangen-Strategien und allgemein verteilten Bedienzeiten erweitert. Diese Warteschlangennetze können offen, geschlossen oder gemischt sein. Kunden können ihre Klassenzugehörigkeit ändern. Es wird gezeigt, daß auch solche Netze, in denen sämtliche Knoten von einem der vier elementaren Wartesystem-Typen

- M/M/c-FCFS,
- M/G/1-PS,
- M/G/ ∞ -IS,
- M/G/1-LCFS PR

sind und sich im statistischen Gleichgewicht befinden, eine Produktformlösung haben. Auf eine weitere Erörterung der sogenannten BCMP-Netze soll an dieser Stelle verzichtet werden, da sie innerhalb dieser Arbeit nicht weiter von Bedeutung sind. Der interessierte Leser sei an [BCMP75] verwiesen.

5. Das Warteschlangenmodell der PAPER-Architektur

Der Modellerstellung kommt bei der analytischen Leistungsbewertung eine zentrale Rolle zu. Das Modell soll die Eigenschaften der PAPER-Architektur so genau wie möglich erfassen, aber auch noch mathematisch handhabbar sein.

Das Leistungsmerkmal für einen Kommunikations-Controller ist die Reaktionszeit auf bestimmte Ereignisse (Ankunft eines Datenpakets, Erhalt eines Kommunikationsauftrags). Diese ist neben der Leistungsfähigkeit der Hardware auch von der Software-Implementierung des Protokolls abhängig. Hardware und Software müssen zum Erreichen einer möglichst kurzen Reaktionszeit aufeinander abgestimmt werden. Es hat sich gezeigt, daß die durch die verwendeten Übertragungsmedien möglichen Übertragungsraten nicht erreicht werden. Dies ist auch auf die zu geringe Leistungsfähigkeit der verwendeten Kommunikations-Controller zurückzuführen.

Das primäre Ziel bei der Leistungsbewertung der PAPER-Architektur ist die Beurteilung des Leistungsverhaltens bei der Parallelverarbeitung eines als PENCIL-Netz formalisierten Kommunikationsprotokolls. Als erster Schritt wird dazu in diesem Kapitel ein Warteschlangenmodell für die PAPER-Architektur konstruiert.

5.1 Die Kunden des Warteschlangenmodells

Die Reaktionszeit auf einzelne Ereignisse setzt sich aus der Abarbeitung mehrerer Teilfunktionen innerhalb des Protokolls zusammen. Bei der PAPER-Architektur ist das Protokoll als PENCIL-Netz formalisiert. Das Abarbeiten der Protokollaktionen ist somit das Schalten von Transitionen bzw. das Ausführen der Transitionsfunktion. Bei Warteschlangenmodellen werden die zu verarbeitenden Elemente als „Kunden“ bezeichnet. Im Warteschlangenmodell der PAPER-Architektur besteht die Kundenmenge \mathcal{K} aus den

- zu verarbeitenden Transitionen des als PENCIL-Netz formalisierten Protokolls (Menge \mathcal{T}),
- Interrupts und anderen externen Ereignissen (im folgenden gemeinsam als *externe Ereignisse* bezeichnet), die Protokollaktionen auslösen (Menge \mathcal{E}).

Die Menge \mathcal{T} ist eine endliche Menge, wogegen die Menge \mathcal{E} eine unendliche Menge ist.

$$\mathcal{K} = \mathcal{T} \cup \mathcal{E}, \quad \mathcal{T} \cap \mathcal{E} = \emptyset$$

Alle Kunden sollen gleich behandelt werden. Die Menge \mathcal{K} besteht also aus nur einer Kundenklasse. Dies gilt bis auf eine Ausnahme: Externe Ereignisse treten nur an den Schaltmaschinen

auf. Sie werden entweder von einer freien Schaltmaschine abgearbeitet oder unterbrechen eine Schaltmaschine, die gerade eine Protokollaktion verarbeitet. Bei der Modellierung muß diese Tatsache durch *statische unterbrechende Prioritäten* berücksichtigt werden. Kunden der Menge \mathcal{E} haben Vorrang vor Kunden der Menge \mathcal{T} .

5.2 Die Konstruktion eines Warteschlangennetzes für die PAPER-Architektur

Bei der Abstraktion der PAPER-Architektur in ein Warteschlangenmodell muß berücksichtigt werden, daß sowohl eine Einzelanalyse der verschiedenen Modellkomponenten als auch eine Analyse des gesamten Modells möglich ist. Dies impliziert eine Modellierung der PAPER-Architektur als *Warteschlangennetz*. Da die Kunden der Menge \mathcal{E} diesem Warteschlangennetz von außen zugeführt werden, muß das Warteschlangennetz *offen* sein.

Zur Modellierung der PAPER-Architektur als Warteschlangennetz werden elementare Wartesysteme mit exponentiell verteilten Zwischenankunfts- und Bedienzeiten verwendet. Die Exponentialverteilung besitzt die Markov-Eigenschaft der Gedächtnislosigkeit. Das Verhalten der zugehörigen Ankunfts- und Bedienprozesse hängt dann nur vom augenblicklichen Zustand ab und ist unabhängig von den vorangegangenen Prozeßzuständen. Diese Modellierungsform bildet einen Kompromiß zwischen der Genauigkeit des Modells und dessen mathematischer Handhabbarkeit. Die Verwendung der Exponentialverteilung erzeugt ein quasi-gleichmäßiges Ankunfts- und Bedienverhalten im Warteschlangenmodell. Das hat zur Folge, daß das Auftreten externer Ereignisse oder die als Bitvektor erfolgende Übergabe von Transitionen in der Kontrolleinheit nicht immer adäquat modelliert wird. Gerade aber wegen der Eigenschaft der Gedächtnislosigkeit und den darauf aufbauenden mathematischen Möglichkeiten haben sich die Markov'schen Wartesysteme bei der analytischen Leistungsbewertung bewährt.

5.2.1 Die Modellierung der Kontrolleinheit

Die Kontrolleinheit der PAPER-Architektur besteht aus drei aktiven Komponenten: Deaktivierer, Auswerter und Aktivierer. Abgearbeitete Protokolltransitionen werden vom Deaktivierer entgegen genommen, der dann die Transitionen wieder frei gibt, die durch das Schalten gesperrt wurden. Durch das Deaktivieren einer Transition wird eine bestimmte Anzahl von Transitionen potentiell schaltfähig. Der Auswerter überprüft diese Transitionen auf ihre Schaltfähigkeit. Im Aktivierer werden die als schaltfähig ermittelten Transitionen auf ihre Aktivierbarkeit geprüft, d.h. ob sie durch andere, im Schalten befindliche Transitionen gesperrt sind.

Der Inzidenzspeicher enthält nur statische Informationen über den Aufbau des als PENCIL-Netz spezifizierten Protokolls. Bei der Modellierung der Kontrolleinheit als Warteschlangennetz kann er vernachlässigt werden. Die Zugriffszeit wird von den Bearbeitungszeiten des Deaktivierers und Aktivierers übernommen. Diese müssen dann berücksichtigen, ob die Zugriffe auf den

auf ihre Schaltfähigkeit zu prüfen ist. Dies wird im Warteschlangennetz durch die Quelle Q_1 modelliert. Sie erzeugt mit der Rate λ_{02} neue Kunden der Menge \mathcal{T} . Die Rate $\lambda_{02} = \lambda_2$ ist daher *direkt* abhängig von der Rate λ_1 . Als schaltfähig erkannte Transitionen werden im Warteschlangennetz als Kunden der Menge \mathcal{T} mit der Rate $p_{23}\lambda_{02}$ der Sperreinheit zugeführt. Nicht aktivierte Transitionen, also schaltfähige aber gesperrte Transitionen, werden durch die Kunden modelliert, die das Warteschlangennetz mit einer Rate von $p'_{30}\lambda_3$ verlassen. Aktivierte Kunden werden der Ausführungseinheit mit der Rate $p_{34}\lambda_3$ zugeführt. Wegen der Local-Balance-Eigenschaft gilt

$$(p'_{30} + p_{34})\lambda_3 = p_{23}\lambda_{02} . \quad (5.1b)$$

Die vom Auswerter als nicht schaltfähig erkannten Transitionen verlassen das Warteschlangensystem mit der Rate $p_{20}\lambda_{02}$.

Die Sperreinheit im Warteschlangennetz erfüllt zwei Funktionen. Sie übernimmt sowohl den Vorgang des Sperrens als auch des Entsperrens von Transitionen. Somit modelliert sie die notwendige Synchronisation bei Zugriffen auf den Sperrvektor. Wenn diese beiden Vorgänge unterschiedliche Bedienzeiten haben, ist dies bei der Modellierung des Bedienprozesses der Sperreinheit durch Mittelwertbildung zu berücksichtigen.

5.2.2 Die Modellierung der Ausführungseinheit

Die Zeitbehaftung der Transitionen und die notwendige Zugriffssynchronisation auf den Markenspeicher machen eine Unterteilung des Schaltvorgangs einer Transition in der PAPER-Architektur in drei Phasen notwendig:

1. Aktivierungsphase

In der Aktivierungsphase werden die zum Schalten benötigten Stelleninhalte aus dem Markenspeicher geholt.

2. Schaltphase

Die Schaltphase umfaßt das eigentliche Schalten (Feuern) der Transition. Die entsprechende Protokollaktion wird durch Abarbeiten der Transitionsfunktion ausgeführt.

3. Deaktivierungsphase

Nach dem Schalten der Transition werden die betroffenen Stelleninhalte im Markenspeicher geändert, wodurch sich der Zustand des PENCIL-Netzes und damit des Protokolls ändert.

Beim Schalten einer Transition erfolgt also *vor* und *nach* dem Schalten ein Zugriff auf den Markenspeicher. Externe Ereignisse werden den Schaltmaschinen über den globalen Bus zugeführt. Deren Abarbeitung besteht nur aus Schalt- und Deaktivierungsphase.

Die Schaltmaschinen lassen sich zunächst durch ein M/M/c-Wartesystem modellieren,

- dem ein M/M/1-Wartesystem als Markenspeicher vor- und nachgelagert ist,

- dem die externen Ereignisse, also Kunden der Menge \mathcal{E} , durch eine Quelle Q_2 zugeführt werden,
- das mit einem weiteren M/M/1-Wartesystem, welches den globalen Speicher modelliert, verbunden ist.

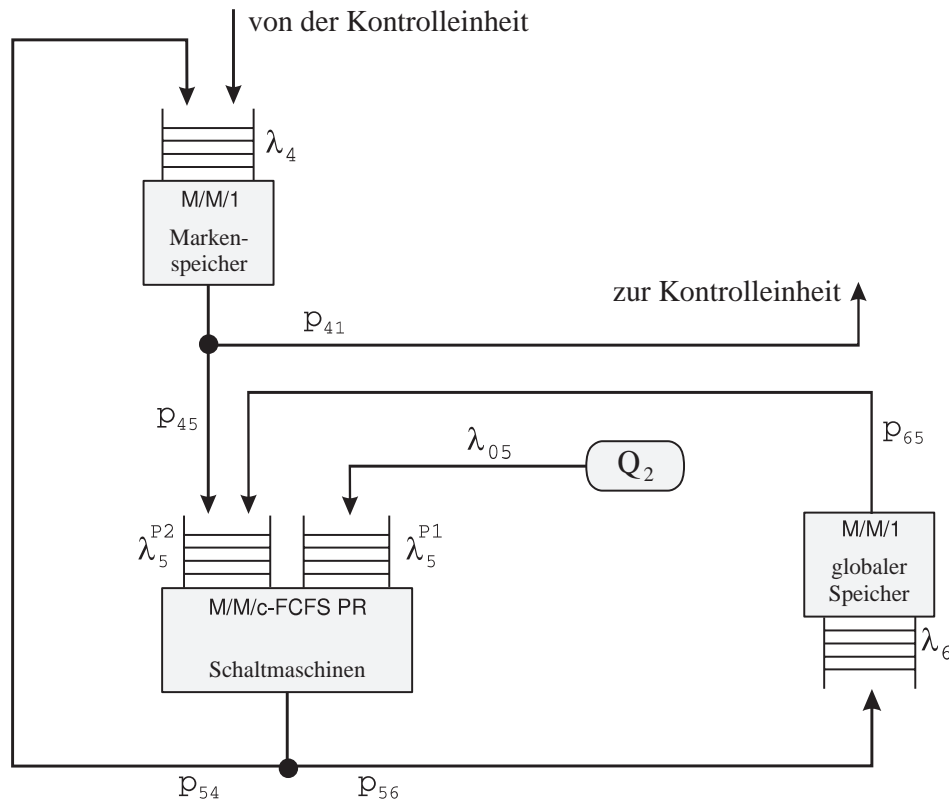


Abbildung 5.2: Das Warteschlangennetz der Ausführungseinheit

Es wurde schon darauf hingewiesen, daß Kunden der Menge \mathcal{E} an den Schaltmaschinen Vorrang vor Kunden der Menge \mathcal{T} haben und dies durch *statische unterbrechende Prioritäten* modelliert wird. Im Abschnitt 4.3.5.1 sind M/M/c-FCFS PR-Wartesysteme vorgestellt worden, deren Zwischenankunfts- und Bedienzeiten exponentiell verteilt sind und die benötigte Bedienstrategie besitzen. Somit werden die Schaltmaschinen als M/M/c-FCFS PR-Wartesystem mit $\mathcal{P} = \{1, 2\}$ modelliert, wobei λ_5^{P1} die Ankunftsrate von Kunden der Menge \mathcal{E} und λ_5^{P2} die Ankunftsrate von Kunden der Menge \mathcal{T} angibt (\Leftrightarrow Abbildung 5.2). Diese Tatsache ist nur für die Einzelanalyse interessant. Wird das gesamte Warteschlangennetz betrachtet, verhalten sich die aggregierten Leistungsgrößen von M/M/c-FCFS PR-Wartesystemen wie die der M/M/c-Wartesysteme (\Leftrightarrow Abschnitt 4.3.5).

Wie im Warteschlangennetz der Kontrolleinheit besitzen auch im Warteschlangennetz der Ausführungseinheit alle Knoten die Local-Balance-Eigenschaft. Das dynamische Ablaufgeschehen wird durch Gesetzmäßigkeiten bei der Transitionsausführung innerhalb der PAPER-Architektur bestimmt.

Die Aktivierungspakete kommen als Kunden aus der Kontrolleinheit am Markenspeicher mit der Rate $p_{34}\lambda_3$ an. Nach ihrer Bedienung werden sie den Schaltmaschinen zugeführt. Dies geschieht

mit der Rate $p_{45}\lambda_4$. Wegen der Local-Balance-Eigenschaft ist

$$p_{45}\lambda_4 = p_{34}\lambda_3 \quad (5.1c)$$

Neben diesen Kunden der Menge \mathcal{T} kommen an den Schaltmaschinen auch Kunden der Menge \mathcal{E} mit der Rate λ_{05} an. Von den in der Ankunftsrate $p_{45}\lambda_4 + \lambda_{05}$ enthaltenen Kunden benutzt ein bestimmter Anteil den globalen Speicher, was durch die Rate $p_{56}\lambda_5$ modelliert wird. Nach dem Speicherzugriff erfolgt die Weiterverarbeitung durch die Schaltmaschinen. Diese Gesetzmäßigkeit ist bei der Modellierung des Bedienprozesses der Schaltmaschinen zu berücksichtigen.

Der Abgangsprozeß der Schaltmaschinen in Richtung Markenspeicher erfolgt mit der Rate $p_{54}\lambda_5$. Diese Kunden gehen nach ihrer Bedienung im Markenspeicher (Deaktivierungsphase) mit einer Rate von $p_{41}\lambda_4$ in die Kontrolleinheit über. Wegen der Local-Balance-Eigenschaft muß

$$p_{41}\lambda_4 = p_{54}\lambda_5 = p_{45}\lambda_4 + \lambda_{05} \stackrel{(5.1c)}{=} p_{34}\lambda_3 + \lambda_{05} \quad (5.1d)$$

gelten. Durch die Gleichung (5.1d) ist zu erkennen, daß die Rate, mit der Kunden von der Kontrolleinheit in der Ausführungseinheit ankommen ($p_{34}\lambda_3$) *kleiner* ist, als die Rate, mit der sie diese verlassen ($p_{34}\lambda_3 + \lambda_{05}$). Es erreichen nur Kunden der Menge \mathcal{T} die Ausführungseinheit. Diese *und* die durch die Quelle Q_2 zugeführten Kunden der Menge \mathcal{E} werden bedient und der Kontrolleinheit zugeführt.

Sind die Bedienzeiten des Markenspeichers während der Aktivierungs- und Deaktivierungsphase verschieden, ist dies bei der Modellierung des Bedienprozesses durch Mittelwertbildung zu berücksichtigen.

5.2.3 Die Modellierung der Verbindung von Kontroll- und Ausführungseinheit

In der PAPER-Architektur sind die Kontroll- und Ausführungseinheit durch zwei FIFO-Warteschlangen miteinander verbunden (\Leftrightarrow Abschnitt 3.2.3). Diese sind in den Warteschlangennetzen von Kontroll- und Ausführungseinheit schon enthalten. Die Verbindungswarteschlange für Kunden, die von der Ausführungseinheit in die Kontrolleinheit übergehen, ist die Warteschlange des Deaktivierers (\Leftrightarrow Abbildung 5.1), deren Ankunftsprozeß die Rate $\lambda_1 = p_{41}\lambda_4$ hat. Für den umgekehrten Fall ist das die Warteschlange des Markenspeichers (\Leftrightarrow Abbildung 5.2). Da dieser sowohl von Kunden der Aktivierungs- als auch der Deaktivierungsphase benutzt wird, sind bei einer Kapazitätsanpassung nur die Kunden der Aktivierungsphase von Bedeutung. Das sind die Kunden mit Ankunftsrate $p_{34}\lambda_3$.

Die Abbildung 5.3 zeigt das gesamte Warteschlangennetz der PAPER-Architektur. Alle Knoten der Knotenmenge $\mathcal{N} = \{1, 2, \dots, 6\}$ sind vom Typ M/M/c, müssen die Local-Balance-Eigenschaft besitzen und haben nur eine Kundenklasse. Durch diese Faktoren ist das Warteschlangennetz der PAPER-Architektur ein *offenes Jackson-Netz*, in dem jeder Knoten isoliert vom gesamten Netz analysiert werden kann (\Leftrightarrow Abschnitt 4.4.2.1).

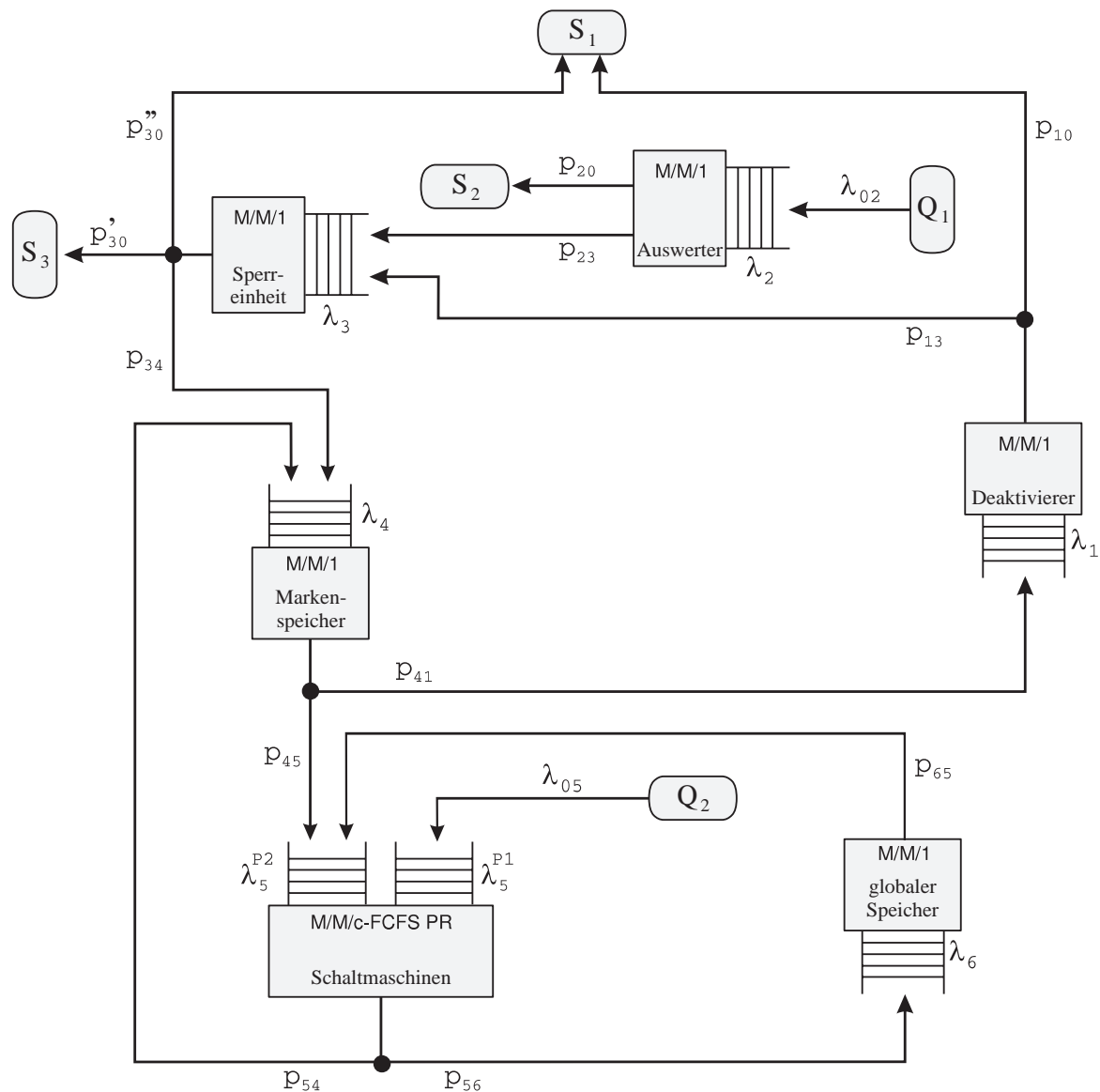


Abbildung 5.3: Das Warteschlangennetz der PAPER-Architektur

5.3 Die Berechnung der Ankunftsrate und Verzweigungswahrscheinlichkeiten

Für das konstruierte Warteschlangennetz der PAPER-Architektur (\clubsuit Abbildung 5.3) erfolgt in diesem Abschnitt die Lastbeschreibung der einzelnen Knoten. Diese umfaßt die Ermittlung der Ankunftsrate und der Verzweigungswahrscheinlichkeiten der einzelnen Knoten. Durch die Ankunftsrate werden die Ankunftsprozesse der Knoten charakterisiert und somit die Kundenlast, die von den einzelnen Knoten zu bewältigen ist. Sie bilden die Grundlage für die Analyse des Warteschlangennetzes.

Nach den im Abschnitt 4.4.1 eingeführten Formalismen für Warteschlangennetze deren Knoten

stationäre Zustände haben, lassen sich die Ankunftsraten durch Lösen des Gleichungssystems

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^6 p_{ji} \cdot \lambda_j, \quad \forall i \in \{1, 2, \dots, 6\} = \mathcal{N} \quad (5.2a)$$

berechnen. Für die Verzweigungswahrscheinlichkeiten muß

$$\sum_{j=0}^6 p_{ij} = 1, \quad \forall i \in \mathcal{N} \quad (5.2b)$$

gelten.

Aus der Abbildung 5.3 können die Gleichungen zur Berechnung von $\vec{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_6)$ direkt aufgestellt werden.

$$\begin{aligned} \lambda_1 &= p_{41} \lambda_4 \\ \lambda_2 &= \lambda_{02} \\ \lambda_3 &= p_{13} \lambda_1 + p_{23} \lambda_2 \\ \lambda_4 &= p_{34} \lambda_3 + p_{54} \lambda_5 \\ \lambda_5 &= \lambda_{05} + p_{45} \lambda_4 + \lambda_6 \\ \lambda_6 &= p_{56} \lambda_5 \end{aligned} \quad (5.3)$$

Die Gesetzmäßigkeiten für den Kundenfluß wurden schon bei der Konstruktion des Warteschlangennetzes beschrieben.

$$p_{13} \lambda_1 \stackrel{(5.1a)}{=} p_{30}'' \lambda_3 \quad (5.4a)$$

$$p_{23} \lambda_2 \stackrel{(5.1b)}{=} p_{23}' \lambda_{02} = (p_{30}' + p_{34}) \lambda_3 \quad (5.4b)$$

$$p_{45} \lambda_4 \stackrel{(5.1c)}{=} p_{34} \lambda_3 \quad (5.4c)$$

$$p_{41} \lambda_4 \stackrel{(5.1d)}{=} p_{54} \lambda_5 = p_{45} \lambda_4 + \lambda_{05} \stackrel{(5.4e)}{=} p_{34} \lambda_3 + \lambda_{05} \quad (5.4d)$$

Es ist $p_{41} \lambda_4 = \lambda_1 \stackrel{(5.2b)}{=} (p_{10} + p_{13}) \lambda_1$. In dem Abgangsprozeß $p_{13} \lambda_1$ sind nur Kunden der Menge \mathcal{T} enthalten. Der Abgangsprozeß $p_{10} \lambda_1$ beinhaltet nur Kunden der Menge \mathcal{E} . Das sind die Kunden, die von der Quelle Q_2 mit einer Rate von λ_{05} dem Warteschlangennetz zugeführt werden. Daher ist

$$\lambda_{05} = p_{10} \lambda_1. \quad (5.4e)$$

Da $\lambda_1 = p_{41} \lambda_4$ ist, gilt mit Gleichung (5.4d)

$$\lambda_1 = p_{54} \lambda_5. \quad (5.4f)$$

Die Gleichung (5.4a) läßt sich auch durch

$$p_{30}'' \lambda_3 = p_{13} \lambda_1 \stackrel{(5.2b)}{=} (1 \Leftrightarrow p_{10}) \lambda_1 \stackrel{(5.4e)}{=} \lambda_1 \Leftrightarrow \lambda_{05} \stackrel{(5.4f)}{=} p_{54} \lambda_5 \Leftrightarrow \lambda_{05} \quad (5.4g)$$

darstellen.

Der weitere Weg zur Berechnung von $\vec{\lambda}$ durch das Gleichungssystem (5.2a) ist von den bekannten Größen abhängig. Im Folgenden soll davon ausgegangen werden, daß

- die Ankunftsrate λ_{05} von Kunden der Menge \mathcal{E} ,
- eine der Verzweigungswahrscheinlichkeiten p_{54} oder p_{56} mit $p_{54} + p_{56} \stackrel{(5.2b)}{=} 1$

bekannt sind. Ebenfalls sei p_{23} oder p_{20} ($p_{20} + p_{23} \stackrel{(5.2b)}{=} 1$) bekannt. p_{23} gibt den Anteil der schaltfähigen Transitionen an den potentiell schaltfähigen Transitionen an, während p_{20} das Gleiche für die als nicht schaltfähig erkannten Transitionen angibt. Dieser Wert läßt sich durch eine Analyse des als PENCIL-Netz formalisierten Protokolls abschätzen.

5.3.1 Die Berechnung von $\vec{\lambda}$ für ein bekanntes λ_{02}

Der Auswerter ermittelt pro Zeiteinheit eine bestimmte Anzahl schaltfähiger Transitionen. Das wird im Warteschlangennetz durch $p_{23}\lambda_2$ modelliert. Die als schaltfähig erkannten Transitionen werden vom Aktivierer auf ihre Aktivierbarkeit geprüft. Nicht aktivierbare Transitionen werden verworfen, was durch die Rate $p'_{30}\lambda_3$ modelliert wird. Dann wird durch

$$\delta_g \stackrel{\text{def.}}{=} \frac{p'_{30}\lambda_3}{p_{23}\lambda_2} = \frac{p'_{30}\lambda_3}{p_{23}\lambda_{02}} \quad (5.5)$$

das Verhältnis von *gesperrten schaltfähigen* zu *schaltfähigen* Transitionen angegeben. Für δ_g gilt $0 \leq \delta_g \leq 1$, denn die in der „Erzeugungsrate“ des Auswerters enthaltenen gesperrten Transitionen kann nicht größer sein als diese selbst. Dementsprechend ist

$$\delta_f = 1 \Leftrightarrow \delta_g, \quad \delta_f, \delta_g \in [0, 1] \quad (5.6)$$

das Verhältnis von *aktivierten* zu *schaltfähigen* Transitionen. Bei der weiteren Ermittlung von $\vec{\lambda}$ wird davon ausgegangen, daß δ_g oder δ_f durch Analysen des zugrunde liegenden PENCIL-Netzes als bekannt vorausgesetzt werden können. Durch δ_g läßt sich das Produkt $p_{34}\lambda_3$ aus der Gleichung (5.4b) bestimmen.

$$\begin{aligned} p_{34}\lambda_3 &= p_{23}\lambda_{02} \Leftrightarrow p'_{30}\lambda_3 \\ &\stackrel{(5.5)}{=} p_{23}\lambda_{02} \Leftrightarrow \delta_g p_{23}\lambda_{02} \\ &= (1 \Leftrightarrow \delta_g) \cdot p_{23}\lambda_2 \\ &\stackrel{(5.6)}{=} \delta_f p_{23}\lambda_{02} \end{aligned}$$

Nach Gleichung (5.4c) gilt dann

$$p_{34}\lambda_3 = p_{45}\lambda_4 = \delta_f p_{23}\lambda_{02} . \quad (5.7)$$

Die in (5.3) aufgestellten Gleichungen zur Berechnung von $\vec{\lambda}$ sind mit den bisher hergeleiteten Gesetzmäßigkeiten wie folgt umformbar.

$$\begin{aligned}
\lambda_1 &= p_{54}\lambda_5 \\
\lambda_2 &= \lambda_{02} \\
\lambda_3 &= p_{23}\lambda_{02} + p_{54}\lambda_5 \Leftrightarrow \lambda_{05} \\
\lambda_4 &= \delta_f p_{23}\lambda_{02} + p_{54}\lambda_5 \\
\lambda_5 &= \lambda_{05} + \delta_f p_{23}\lambda_{02} + \lambda_6 \\
\lambda_6 &= p_{56}\lambda_5
\end{aligned} \tag{5.8}$$

Dann sind nur noch die zu errechnenden λ_i unbekannt.

Durch Lösen des linearen Gleichungssystems

$$\begin{aligned}
\lambda_1 & \Leftrightarrow p_{54}\lambda_5 & = & 0 \\
\lambda_2 & & = & \lambda_{02} \\
\lambda_3 & \Leftrightarrow p_{54}\lambda_5 & = & p_{23}\lambda_{02} \Leftrightarrow \lambda_{05} \\
\lambda_4 & \Leftrightarrow p_{54}\lambda_5 & = & \delta_f p_{23}\lambda_{02} \\
\lambda_5 & \Leftrightarrow \lambda_6 & = & \lambda_{05} + \delta_f p_{23}\lambda_{02} \\
& \Leftrightarrow p_{56}\lambda_5 & \lambda_6 & = 0
\end{aligned} \tag{5.9}$$

nach dem Gauß'schen Lösungsverfahren erhält man die gesuchten Ankunftsraten λ_i .

$$\begin{aligned}
\lambda_1 &= \lambda_{05} + \delta_f p_{23}\lambda_{02} \\
\lambda_2 &= \lambda_{02} \\
\lambda_3 &= (1 + \delta_f) \cdot p_{23}\lambda_{02} \\
\lambda_4 &= \lambda_{05} + 2 \cdot \delta_f p_{23}\lambda_{02} \\
\lambda_5 &= (\lambda_{05} + \delta_f p_{23}\lambda_{02}) \cdot \frac{1}{p_{54}} \\
\lambda_6 &= (\lambda_{05} + \delta_f p_{23}\lambda_{02}) \cdot \frac{p_{56}}{p_{54}}
\end{aligned} \tag{5.10}$$

λ_{05} , λ_{02} , δ_f , p_{23} , p_{54} und p_{56} sind positive Werte. Dann ist $\vec{\lambda} > \vec{0}$ und somit eine gültige Lösung für das Gleichungssystem (5.9).

Mit den errechneten λ_i lassen sich die durch den Kundenfluß bestimmten Verzweigungswahrscheinlichkeiten für das Warteschlangennetz berechnen.

$$\begin{aligned}
p_{10} &\stackrel{(5.4e)}{=} \frac{\lambda_{05}}{\lambda_1} = \frac{\lambda_{05}}{\lambda_{05} + \delta_f p_{23} \lambda_{02}} \\
p_{13} &\stackrel{(5.2b)}{=} 1 \Leftrightarrow p_{10} = \frac{\delta_f p_{23} \lambda_{02}}{\lambda_{05} + \delta_f p_{23} \lambda_{02}} \\
p'_{30} &\stackrel{(5.5)}{=} \frac{\delta_g p_{23} \lambda_{02}}{\lambda_3} = \frac{\delta_g}{1 + \delta_f} \\
p''_{30} &\stackrel{(5.4a)}{=} \frac{p_{13} \lambda_1}{\lambda_3} = \frac{\delta_f}{1 + \delta_f} \\
p_{34} &\stackrel{(5.7)}{=} \frac{\delta_f p_{23} \lambda_{02}}{\lambda_3} = \frac{\delta_f}{1 + \delta_f} \\
p_{41} &\stackrel{(5.3)}{=} \frac{\lambda_1}{\lambda_4} = \frac{\lambda_{05} + \delta_f p_{23} \lambda_{02}}{\lambda_{05} + 2 \cdot \delta_f p_{23} \lambda_{02}} \\
p_{45} &\stackrel{(5.7)}{=} \frac{\delta_f p_{23} \lambda_{02}}{\lambda_4} = \frac{\delta_f p_{23} \lambda_{02}}{\lambda_{05} + 2 \cdot \delta_f p_{23} \lambda_{02}}
\end{aligned} \tag{5.11}$$

Wie durch die Gleichungen (5.10) zu erkennen ist, sind die Ankunftsrate λ_i im wesentlichen von drei Faktoren abhängig:

- λ_{05} Durch die Rate λ_{05} werden dem Warteschlangennetz Kunden der Menge \mathcal{E} zugeführt.
- λ_{02} Die Rate λ_{02} erzeugt am Auswerter neue Kunden der Menge \mathcal{T} .
- $\delta_f p_{23}$ Das Produkt $\delta_f p_{23}$ gibt den Anteil der durch die Quelle Q_2 erzeugten Kunden an, die *nicht* von einer der Senken S_2 oder S_3 aufgenommen werden.

Die von der PAPER-Architektur auszuführenden Protokollaktionen sind in erster Linie vom Auftreten externer Ereignisse abhängig. Durch sie werden die Protokollaktionen initiiert. Welche und wieviele Protokollaktionen auszuführen sind, ist von der Struktur des PENCIL-Netzes und der im Schalten befindlichen Transitionen abhängig.

Wenn *keine* externen Ereignisse auftreten ($\lambda_{05} = 0$), werden nur Protokollaktionen ausgeführt, die unabhängig von externen Ereignissen sind. Werden dem Auswerter keine Transitionen zur Prüfung der Schaltfähigkeit übergeben ($\lambda_{02} = 0$) oder findet er keine schaltfähigen Transitionen ($p_{23} = 0$), dann werden nur die externen Ereignisse an den Schaltmaschinen bearbeitet aber das Protokoll reagiert auf diese nicht. Ein ähnlicher Fall tritt ein, wenn alle als schaltfähig erkannten Transitionen gesperrt sind ($\delta_g = 1, \delta_f = 0$). Die kürzeste Reaktionszeit auf externe Ereignisse wird erreicht, wenn alle als schaltfähig erkannten Transitionen aktiviert werden können ($\delta_f = 1, \delta_g = 0$).

5.3.2 Die Berechnung von $\vec{\lambda}$ für $\lambda_{02} = \gamma\lambda_1$

Bei der Modellierung der Kontrolleinheit (\diamond Abschnitt 5.2.1) wurde darauf hingewiesen, daß die Ankunftsrate $\lambda_2 = \lambda_{02}$ direkt von der Rate λ_1 abhängig ist.

$$\lambda_2 = \lambda_{02} = \gamma\lambda_1, \quad \gamma \geq 0 \quad (5.12)$$

Jeder Kunde, der im Deaktivierer bedient wird, erzeugt γ neue Kunden. Das sind die inzidenz-abhängigen, potentiell schaltfähigen Transitionen, die vom Auswerter auf ihre Schaltfähigkeit zu prüfen sind. Das Verhältnis von *gesperrten schaltfähigen* zu *schaltfähigen* Transitionen ist nun durch

$$\delta_g = \frac{p'_{30}\lambda_3}{p_{23}\gamma\lambda_1}$$

gegeben.

Mit den Gleichungen (5.8) und den Gesetzmäßigkeiten für den Kundenfluß im Warteschlangennetz der PAPER-Architektur, lassen sich die Gleichungen zur Berechnung von $\vec{\lambda}$ für $\lambda_{02} = \gamma\lambda_1$ durch

$$\begin{aligned} \lambda_1 &= p_{54}\lambda_5 \\ \lambda_2 &= \gamma p_{54}\lambda_5 \\ \lambda_3 &= (1 + p_{23}\gamma) \cdot p_{54}\lambda_5 \Leftrightarrow \lambda_{05} \\ \lambda_4 &= (1 + \delta_f p_{23}\gamma) \cdot p_{54}\lambda_5 \\ \lambda_5 &= \frac{1}{1 \Leftrightarrow \delta_f p_{23}\gamma p_{54}} \cdot (\lambda_{05} + \lambda_6) \\ \lambda_6 &= p_{56}\lambda_5 \end{aligned} \quad (5.13)$$

aufstellen.

Der Lösungsvektor $\vec{\lambda}$ für das lineare Gleichungssystem

$$\begin{array}{rcll} \lambda_1 & \Leftrightarrow p_{54}\lambda_5 & = & 0 \\ \lambda_2 & \Leftrightarrow \gamma p_{54}\lambda_5 & = & 0 \\ \lambda_3 & \Leftrightarrow (1 + p_{23}\gamma)p_{54}\lambda_5 & = & \Leftrightarrow \lambda_{05} \\ \lambda_4 & \Leftrightarrow (1 + \delta_f p_{23}\gamma)p_{54}\lambda_5 & = & 0 \\ \lambda_5 & \Leftrightarrow \frac{1}{1 - \delta_f p_{23}\gamma p_{54}} \lambda_6 & = & \frac{1}{1 - \delta_f p_{23}\gamma p_{54}} \lambda_{05} \\ & \Leftrightarrow p_{56}\lambda_5 & \Leftrightarrow \lambda_6 & = 0 \end{array} \quad (5.14)$$

beinhaltet die gesuchten Ankunftsraten λ_i .

$$\begin{aligned}
\lambda_1 &= \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05} \\
\lambda_2 &= \frac{\gamma}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05} \\
\lambda_3 &= \frac{(1 + \delta_f) \cdot p_{23} \gamma}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05} \\
\lambda_4 &= \frac{1 + \delta_f p_{23} \gamma}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05} \\
\lambda_5 &= \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \frac{1}{p_{54}} \lambda_{05} \\
\lambda_6 &= \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \frac{p_{56}}{p_{54}} \lambda_{05}
\end{aligned} \tag{5.15}$$

λ_{05} , δ_f , p_{23} und γ sind positive Werte. Damit nun $\vec{\lambda} > \vec{0}$ ist, muß

$$1 \Leftrightarrow \delta_f p_{23} \gamma > 0 \quad \Leftrightarrow \quad \delta_f p_{23} \gamma < 1 \tag{5.16}$$

gelten. Diese Restriktion läßt die Frage aufkommen, ob das Warteschlangennetz die PAPER-Architektur korrekt modelliert. Im Abschnitt 5.3.1 wurde eine gültige Lösung für $\vec{\lambda}$ hergeleitet, bei der λ_{02} als bekannt vorausgesetzt wird. Das führt zu der Vermutung, daß die Restriktion (5.16) durch den Kundenfluß und die Abhängigkeit von λ_{02} und λ_1 hervorgerufen wird.

Durch das Deaktivieren einer Transition t werden in der PAPER-Architektur inzidenzbedingt im Mittel γ Transitionen potentiell schaltfähig. Der Auswerter erkennt

$$t_s \stackrel{\text{def.}}{=} p_{23} \gamma \tag{5.17}$$

Transitionen, die durch das Schalten der Transition t schaltfähig geworden sind. Es ist $t_s \leq \gamma$, da für p_{23} als Verzweigungswahrscheinlichkeit $p_{23} \in [0, 1]$ gilt. Von diesen t_s schaltfähigen Transitionen können nur

$$t_a \stackrel{\text{def.}}{=} \delta_f p_{23} \gamma \tag{5.18}$$

Transitionen aktiviert werden. Die restlichen $\delta_g p_{23} \gamma = (1 \Leftrightarrow \delta_f) p_{23} \gamma$ Transitionen sind durch im Schalten befindliche Transitionen gesperrt. Da $\delta_f \in [0, 1]$, ist $t_a \leq t_s \leq \gamma$. Die Abbildung 5.4 verdeutlicht diesen Zusammenhang.

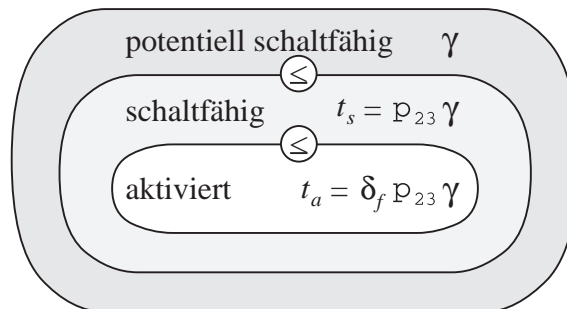


Abbildung 5.4: Zusammenhang zwischen *potentiell schaltfähig*, *schaltfähig* und *aktiviert*

Somit hat das *Deaktivieren* einer Transition *das Aktivieren* von im Mittel $t_a = \delta_f p_{23} \gamma$ Transitionen

zur Folge und es muß nach (5.16)

$$0 < 1 \Leftrightarrow \delta_f p_{23} \gamma = 1 \Leftrightarrow t_a < 1 \quad \Leftrightarrow \quad t_a < 1$$

gelten. Im Mittel dürfen pro deaktivierter Transition *weniger als eine* Transition aktiviert werden.

Erweitert man die Restriktion (5.16) mit λ_1 , so gilt:

$$\begin{aligned} & \delta_f p_{23} \gamma < 1 \\ \Leftrightarrow & \delta_f p_{23} \gamma \lambda_1 < \lambda_1 \\ \stackrel{(5.12)}{\Leftrightarrow} & \delta_f p_{23} \lambda_2 < \lambda_1 \\ \stackrel{(5.7)}{\Leftrightarrow} & p_{34} \lambda_3 < \lambda_1 \stackrel{(5.3)}{=} p_{41} \lambda_4 \end{aligned} \quad (5.19)$$

Die Rate, mit der Kunden im Warteschlangennetz von der Sperreinheit zum Markenspeicher übergehen ($p_{34} \lambda_3$), muß im Mittel kleiner sein als die Rate, mit der sie den Markenspeicher in Richtung Deaktivierer verlassen ($p_{41} \lambda_4$).

Übertragen auf die PAPER-Architektur bedeutet das, daß pro Zeiteinheit im Mittel weniger Transitionen von der Kontrolleinheit in die Ausführungseinheit gelangen dürfen als umgekehrt. Das ist der Fall, da neben den abzuarbeitenden Transitionen, die von der Kontrolleinheit in die Ausführungseinheit übergegangen sind, auch die externen Ereignisse nach ihrer Bearbeitung an die Kontrolleinheit weitergeleitet werden¹.

Die in Ungleichung (5.16) aufgestellte Restriktion bedeutet für die Modellierung der PAPER-Architektur durch das im Abschnitt 5.2 konstruierte Warteschlangennetz keine Einschränkung, sondern wird durch Gesetzmäßigkeiten bei der Abarbeitung von Protokollaktionen und der Notwendigkeit der Local-Balance-Eigenschaft hervorgerufen.

Mit den durch die Gleichungen (5.15) errechneten λ_i lassen sich die durch den Kundenfluß bestimmten Verzweigungswahrscheinlichkeiten wie in den Gleichungen (5.11) für das Warteschlangennetz bestimmen.

$$\begin{aligned} p_{10} &= 1 \Leftrightarrow \delta_f p_{23} \gamma & p_{34} &= \frac{\delta_f}{1 + \delta_f} \\ p_{13} &= \delta_f p_{23} \gamma & p_{41} &= \frac{1}{1 + \delta_f p_{23} \gamma} \\ p'_{30} &= \frac{\delta_g}{1 + \delta_f} & p_{45} &= \frac{\delta_f p_{23} \gamma}{1 + \delta_f p_{23} \gamma} \\ p''_{30} &= \frac{\delta_f}{1 + \delta_f} & & \end{aligned} \quad (5.20)$$

¹ ⇨ Bemerkung zur Gleichung (5.1d) auf Seite 62

5.4 Die Berechnung der Verzweigungswahrscheinlichkeiten p_{54} und p_{56}

Bei der Berechnung der Ankunftsrate für das Warteschlangennetz der PAPER-Architektur wurde u.a. davon ausgegangen, daß die Verzweigungswahrscheinlichkeiten p_{54} und p_{56} bekannt sind. Die Werte dieser Verzweigungswahrscheinlichkeiten sind von der Anzahl der Transitionen abhängig, die während ihrer Bearbeitung durch die Schaltmaschinen auf den globalen Speicher der PAPER-Architektur zugreifen.

Die Schaltmaschinen wurden als ein M/M/c-FCFS PR-Wartesystem modelliert. Dieses besitzt nach Abschnitt 4.3.5 die Eigenschaft, daß sich die aggregierten Leistungsgrößen wie die eines M/M/c-Wartesystems verhalten. Die Schaltmaschinen haben einen Poisson'schen Ankunftsprozeß mit der aggregierten Ankunftsrate λ_5 . In dieser Rate sind auch die Kunden enthalten, die schon eine Teilbedienung durch die Schaltmaschinen erhalten haben. Diese kommen an den Schaltmaschinen mit einer Rate von $p_{65}\lambda_6 = \lambda_6$ an. Wegen der Verschmelzungseigenschaft des Poisson-Prozesses (\Leftrightarrow Abschnitt 4.2.3.5) kommen „neue“ Kunden mit der Rate $\lambda_5 \Leftrightarrow \lambda_6$ an den Schaltmaschinen an.

$$\begin{aligned} \lambda_5 \Leftrightarrow \lambda_6 &\stackrel{(5.3)}{=} p_{45}\lambda_4 + \lambda_{05} \\ &\stackrel{(5.4d)}{=} p_{41}\lambda_4 \\ &\stackrel{(5.3)}{=} \lambda_1 \end{aligned}$$

Diese ankommenden Kunden greifen mit der Wahrscheinlichkeit p_{glob} während der Bedienung durch die Schaltmaschinen auf den globalen Speicher zu. Somit besitzt der Abgangsprozeß der Schaltmaschinen in Richtung globalen Speicher die Rate

$$p_{56}\lambda_5 = p_{glob}\lambda_1 .$$

Daraus läßt sich nun p_{56} berechnen.

$$\begin{aligned} p_{56}\lambda_5 &= p_{glob}\lambda_1 \\ \Leftrightarrow p_{56} &= p_{glob} \cdot \frac{\lambda_1}{\lambda_5} \\ &\stackrel{(5.15)}{=} p_{glob} \cdot \frac{\frac{1}{1-\delta_f p_{23}\gamma} \lambda_{05}}{\frac{1}{1-\delta_f p_{23}\gamma} \frac{1}{p_{54}} \lambda_{05}} \\ &= p_{glob} \cdot p_{54} \\ &\stackrel{(5.2b)}{=} p_{glob} \cdot (1 \Leftrightarrow p_{56}) \\ \Leftrightarrow p_{56} &= \frac{p_{glob}}{1 + p_{glob}} \end{aligned} \tag{5.21}$$

Da $p_{54} + p_{56} = 1$ ist, ist p_{54} direkt berechenbar.

$$\begin{aligned} p_{54} &= 1 \Leftrightarrow p_{56} \\ &\stackrel{(5.21)}{=} \frac{1}{1 + p_{glob}} \end{aligned} \tag{5.22}$$

Die einzigen Ankunftsrate, die direkt von p_{54} und p_{56} abhängen, sind λ_5 und λ_6 . Mit den hergeleiteten Berechnungsvorschriften für p_{54} und p_{56} gilt dann:

$$\begin{aligned}
 \lambda_5 &\stackrel{(5.10)}{=} (\lambda_{05} + \delta_f p_{23} \lambda_{02}) \cdot \frac{1}{p_{54}} & \lambda_6 &\stackrel{(5.10)}{=} (\lambda_{05} + \delta_f p_{23} \lambda_{02}) \cdot \frac{p_{56}}{p_{54}} \\
 &= (\lambda_{05} + \delta_f p_{23} \lambda_{02}) \cdot (1 + p_{glob}) & &= (\lambda_{05} + \delta_f p_{23} \lambda_{02}) \cdot p_{glob} \\
 \lambda_5 &\stackrel{(5.15)}{=} \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \frac{1}{p_{54}} \lambda_{05} & \lambda_6 &\stackrel{(5.15)}{=} \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \frac{p_{56}}{p_{54}} \lambda_{05} & (5.23) \\
 &= \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \cdot (1 + p_{glob}) \cdot \lambda_{05} & &= \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \cdot p_{glob} \cdot \lambda_{05}
 \end{aligned}$$

Auf die Bedeutung der hier berechneten Verzweigungswahrscheinlichkeiten wird bei der Analyse der Ankunftsprozesse im Warteschlangennetz der PAPER-Architektur näher eingegangen.

In diesem Kapitel wurden die ersten Schritte für die analytische Leistungsbewertung der PAPER-Architektur durchgeführt. Es wurde ein Warteschlangenmodell konstruiert und auf dessen Basis die Berechnungsvorschriften für die Lastmodellierung und die Interaktionen der einzelnen Knoten hergeleitet. Diese werden im Anhang B noch einmal zusammenfassend aufgelistet.

Auf der Basis der Ankunftsrate und Verzweigungswahrscheinlichkeiten wird in den nächsten Kapiteln das Verhalten des Warteschlangennetzes untersucht, um daraus Rückschlüsse auf die Leistungsfähigkeit der PAPER-Architektur ziehen zu können.

6. Die Analyse des Warteschlangenmodells

6.1 Die Analyse der Ankunftsraten

Im Abschnitt 5.3 wurden Berechnungsvorschriften für die Ankunftsraten und Verzweigungswahrscheinlichkeiten des Warteschlangennetzes der PAPER-Architektur hergeleitet. Die Ankunftsraten charakterisieren die Poisson'schen Ankunftsprozesse der Knoten und somit die Kundenlast, die von den einzelnen Knoten zu bewältigen ist. Diese werden durch den Faktor

$$F \stackrel{\text{def.}}{=} \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \stackrel{(5.18)}{=} \frac{1}{1 \Leftrightarrow t_a}, \quad t_a \in [0, 1) \quad (6.1)$$

und die Rate λ_{05} , mit der dem Warteschlangennetz Kunden der Menge \mathcal{E} zugeführt werden, bestimmt¹.

Die zentrale Rolle der externen Ankunftsrate λ_{05} für die Ankunftsraten aller Knoten des Warteschlangennetzes ist durch die Konstruktion der PAPER-Architektur begründet. Erst durch das Auftreten von externen Ereignissen wird das Ausführen von Protokollaktionen ausgelöst.

Der Wert von F ist von t_a abhängig, das nach Gleichung (5.18) das Produkt der folgenden drei Faktoren ist:

- γ Eine inzidenzabhängige Anzahl von Transitionen, die nach dem Schalten einer Transition oder dem Auftreten eines externen Ereignisses potentiell schaltfähig geworden sind.
- p_{23} Der Anteil von γ , der vom Auswerter als schaltfähig erkannt wird.
- δ_f Das Verhältnis von aktivierten zu schaltfähigen Transitionen.

Durch t_a wird die mittlere Anzahl von Kunden der Menge \mathcal{T} angegeben, die als Folge der Bearbeitung eines Kunden der Menge \mathcal{K} durch den Deaktivierer, von der Sperreinheit an die Ausführungseinheit weitergeleitet werden. Somit gibt t_a die mittlere Anzahl von Transitionen an, die nach dem Schalten einer Transition aktiviert sind: Ist in der PAPER-Architektur eine Transition oder ein externes Ereignis bearbeitet worden, so können anschließend im Mittel t_a andere Transitionen ausgeführt werden.

Die Bedeutung von F soll durch zwei Extremfälle untersucht werden.

¹ Da $t_a = \delta_f p_{23} \gamma \in [0, 1)$ ist, ist F der Grenzwert der geometrischen Reihe $\sum_{i=0}^{\infty} (t_a)^i$.

$$1. t_a = 0 \Rightarrow F = 1$$

Die Ankunftsrate im Warteschlangennetz werden nur durch Kunden der Menge \mathcal{E} bestimmt, die dem Warteschlangennetz durch die Quelle Q_2 mit der Rate λ_{05} zugeführt werden. In einer solchen Situation bearbeitet die PAPER-Architektur nur die externen Ereignisse, aber es erfolgt keine Reaktion durch das Protokoll. Die Ursache für ein solches Verhalten liegt in einem der drei folgenden Fälle:

- a) Es existieren keine potentiell schaltfähigen Transitionen ($\gamma = 0$).
- b) Der Auswerter kann keine schaltfähigen Transitionen ermitteln ($p_{23} = 0, p_{20} = 1$).
- c) Alle als schaltfähig erkannten Transitionen sind gesperrt ($\delta_f = 0, \delta_g = 1$).

$$2. t_a \rightarrow 1 \Rightarrow F \rightarrow \infty$$

Das Bedienen eines Kunden der Menge \mathcal{K} durch den Deaktivierer hat zur Folge, daß die Sperreinheit zu *fast* jedem dieser Kunden einen Kunden der Menge \mathcal{T} an die Ausführungseinheit weiterleitet. Eine solche Situation wird durch einen der drei folgenden Fälle hervorgerufen:

- a) Fast jede potentiell schaltfähige Transition wird als schaltfähig erkannt und aktiviert.
- b) Es existieren viele potentiell schaltfähige Transitionen, von denen aber nur annähernd $\delta_f p_{23} = \frac{1}{\gamma}$ Transitionen aktiviert werden können.
- c) Erkennt der Auswerter viele der potentiell schaltfähigen Transitionen als schaltfähig, so ist das Verhältnis von aktivierten zu schaltfähigen Transitionen annähernd $\delta_f = \frac{1}{p_{23}\gamma}$.

Die Abbildung 6.1 zeigt den Verlauf des Faktors F als Funktion von t_a .

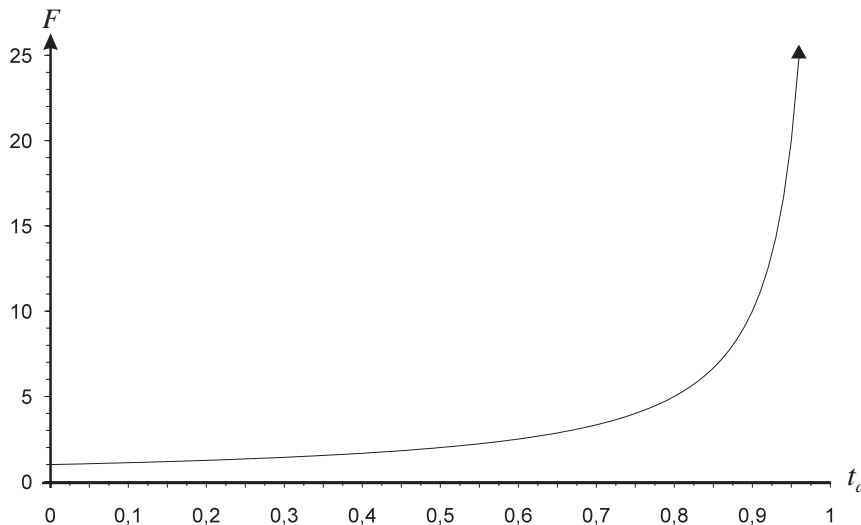


Abbildung 6.1: Der Faktor F als Funktion von $t_a, t_a \in [0, 1)$

Die Betrachtungen des Faktors F lassen Aussagen für das Verhalten der PAPER-Architektur im Hinblick auf die Struktur des PENCIL-Netzes zu.

Besitzt das PENCIL-Netz eine Vielzahl von unabhängigen Transitionssequenzen, so wird je

Transitionssequenz zwar nur eine Transition als potentiell schaltfähig erkannt, die, wenn ihre Schaltbedingung erfüllt ist, auch aktiviert wird. Im umgekehrten Fall werden viele Transitionen als potentiell schaltfähig erkannt. Aber wegen des hohen Konnektivitätsgrades sind viele dieser Transitionen durch in Bearbeitung befindliche Transitionen gesperrt. Bei der Umsetzung eines Protokolls in ein PENCIL-Netz ist darauf zu achten, daß die Abhängigkeit von Transitionen untereinander auf ein Mindestmaß reduziert wird. Dann können viele der als schaltfähig erkannten Transitionen aktiviert werden. Durch die unabhängigen Transitionssequenzen wird ein hohes Maß an Parallelität erreicht, das einer entsprechend schnellen Abarbeitung von Protokollaktionen gleichkommt.

Die Ankunftsraten λ_i aus den Gleichungen (5.15) lassen sich durch

$$\lambda_i = F \cdot f_i \cdot \lambda_{05} , \quad \forall i \in \mathcal{N} \quad (6.2)$$

ausdrücken, wobei f_i ein knotenspezifischer Faktor ist.

Die Gleichung (6.2) läßt erkennen, daß sich die Ankunftsraten linear mit der Ankunftsrate λ_{05} ändern. Verändert sich λ_{05} um den Faktor k , dann verändern sich alle λ_i um diesen Faktor. Die Gesetzmäßigkeiten im Kundenfluß zwischen den Knoten bleiben davon unberührt, da die Verzweigungswahrscheinlichkeiten unabhängig von λ_{05} sind. Die zentrale Rolle der Ankunftsrate λ_{05} für die Ankunftsrate λ_i aller Knoten des Warteschlangennetzes liegt in der Bedeutung der externen Ereignisse für die PAPER-Architektur. Erst durch sie werden die Protokollaktionen in Form von Transitionen des PENCIL-Netzes ausgelöst. Treten also viele externe Ereignisse ein, so erhöht sich die Anzahl der auszuführenden Transitionen, was sich im Warteschlangennetz in höheren Ankunftsraten ausdrückt.

Änderungen des Faktors F werden durch t_a hervorgerufen und sind von der Struktur des PENCIL-Netzes abhängig. Das Verhalten von F in Bezug auf t_a spiegelt sich dann in der Höhe der Ankunftsrate wieder. Dieser Sachverhalt wird in der Abbildung 6.2 für einige ausgewählte Werte von $t_a \in [0, 1)$ dargestellt. Während die externen Ereignisse das Transitionsaufkommen in der PAPER-Architektur von *außen* beeinflussen, werden durch t_a die Ankunftsrate im Warteschlangennetz von *innen*, d.h. durch die Anzahl der aktivierten Transitionen, verändert.

Aus den Berechnungsvorschriften der Ankunftsrate in den Gleichungen (5.15) läßt sich erkennen, daß der Auswerter und die Sperreinheit die höchsten Ankunftsrate und damit das höchste Lastaufkommen haben. In der PAPER-Architektur sind der Auswerter und der Aktivierer für die Bereitstellung neuer Transitionen verantwortlich. Sie müssen aus einer Menge von Transitionen diejenigen herausfiltern, die dann in der Ausführungseinheit zur Bearbeitung kommen. Für die Leistungsfähigkeit der PAPER-Architektur kommt ihnen somit eine besondere Bedeutung zu.

6.2 Die Analyse der Verzweigungswahrscheinlichkeiten

Auch einige Verzweigungswahrscheinlichkeiten ändern sich bei Änderungen von t_a . Durch die Verzweigungswahrscheinlichkeiten sind die Gesetzmäßigkeiten für den Kundenfluß im Warte-

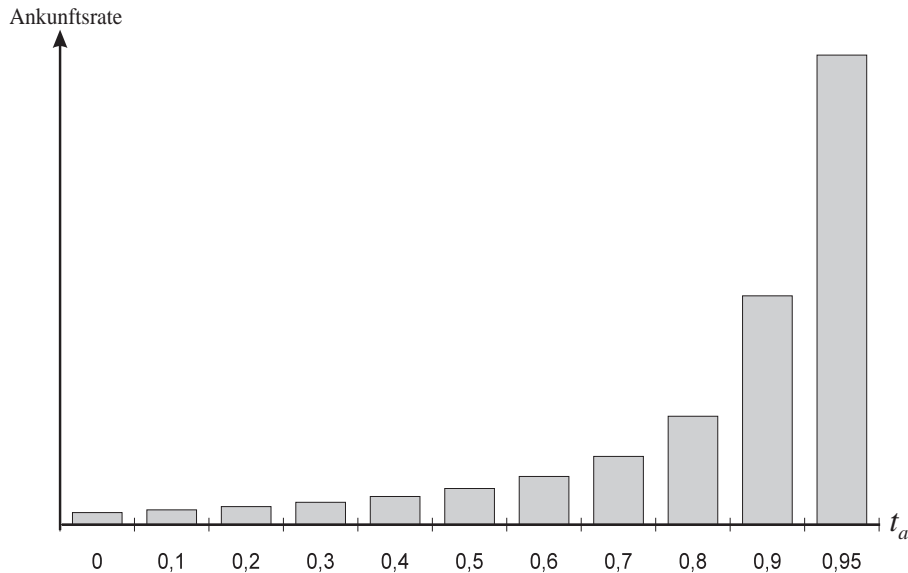


Abbildung 6.2: Das Verhalten der Ankunftsrate für einige ausgewählte Werte von t_a

schlangennetz modelliert worden. In erster Linie sind davon die Verzweigungswahrscheinlichkeiten des Markenspeichers und des Deaktivierers betroffen.

6.2.1 Die Verzweigungswahrscheinlichkeiten des Markenspeichers

Für die Verzweigungswahrscheinlichkeiten des Markenspeichers gilt:

$$p_{41} \stackrel{(5.20)}{=} \frac{1}{1 + \delta_f p_{23} \gamma} \stackrel{(5.18)}{=} \frac{1}{1 + t_a} \quad p_{45} \stackrel{(5.20)}{=} \frac{\delta_f p_{23} \gamma}{1 + \delta_f p_{23} \gamma} \stackrel{(5.18)}{=} \frac{t_a}{1 + t_a}$$

Sie haben den in Abbildung 6.3 dargestellten Verlauf für ein sich änderndes t_a .

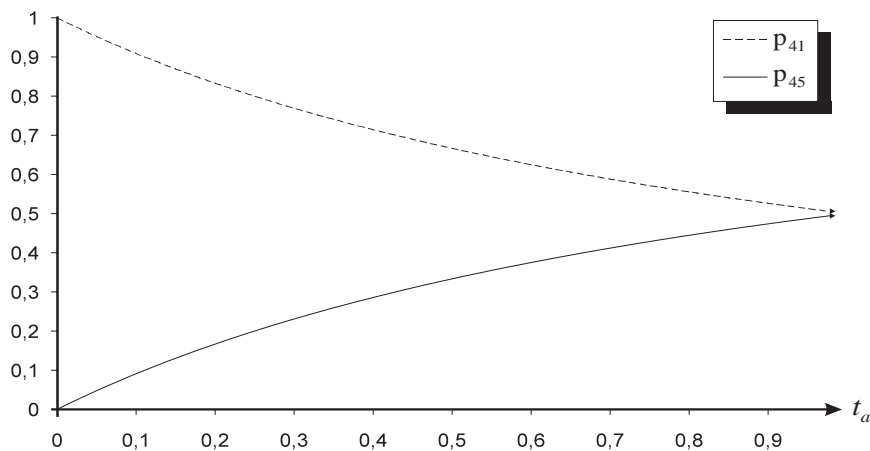


Abbildung 6.3: Das Verhalten der Verzweigungswahrscheinlichkeiten p_{41} und p_{45} bei Änderungen von t_a , $t_a \in [0, 1)$

Für $t_a \rightarrow 1$ nähern sich p_{41} und p_{45} einem Wert von 0,5 an. Es gilt $p_{45} \lambda_4 \stackrel{(5.7)}{=} p_{34} \lambda_3$ und $p_{34} \lambda_3 \stackrel{(5.19)}{<} p_{41} \lambda_4$. Die Rate, mit der Kunden von der Kontrolleinheit in die Ausführungsein-

heit übergehen ($p_{34}\lambda_3$), ist *fast* gleich der Rate, mit der sie von der Ausführungseinheit zu Kontrolleinheit übergehen ($p_{41}\lambda_4$). Diese *annähernde* Gleichheit kommt daher, daß in der Rate, mit der Kunden von der Ausführungseinheit zur Kontrolleinheit übergehen, die durch die Quelle Q_2 erzeugten neuen Kunden enthalten sind. Der Anteil der durch die Quelle Q_2 zugeführten Kunden der Menge \mathcal{E} an der Rate $p_{41}\lambda_4$ ist sehr gering, da die Raten $p_{41}\lambda_4$ und $p_{45}\lambda_4$ wegen ihrer annähernden Gleichheit von p_{41} und p_{45} hauptsächlich durch $p_{34}\lambda_3$ bestimmt sind. Es sind somit erheblich mehr Kunden der Menge \mathcal{T} als Kunden der Menge \mathcal{E} im Warteschlangennetz.

In der PAPER-Architektur sind in einer solchen Situation hauptsächlich Transitionen des PENCIL-Netzes im Umlauf. Die externen Ereignisse spielen wegen ihrer geringen Anzahl im Vergleich zu den Transitionen eine untergeordnete Rolle.

6.2.2 Die Verzweigungswahrscheinlichkeiten des Deaktivierers

Die Verzweigungswahrscheinlichkeiten des Deaktivierers lassen sich durch

$$p_{10} \stackrel{(5.20)}{=} 1 \Leftrightarrow \delta_f p_{23} \gamma \stackrel{(5.18)}{=} 1 \Leftrightarrow t_a \qquad p_{13} \stackrel{(5.20)}{=} \delta_f p_{23} \gamma \stackrel{(5.18)}{=} t_a$$

berechnen. Für ein sich änderndes t_a haben sie den in Abbildung 6.4 dargestellten linearen Verlauf.

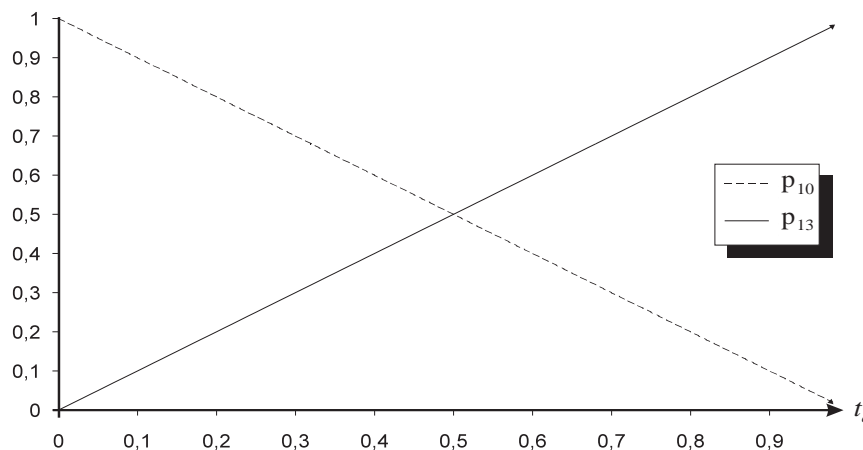


Abbildung 6.4: Das Verhalten der Verzweigungswahrscheinlichkeiten p_{10} und p_{13} bei Änderungen von t_a , $t_a \in [0, 1)$

Für $t_a = 0.5$ sind die Verzweigungswahrscheinlichkeiten des Deaktivierers gleich. In diesem Fall verlassen pro Zeiteinheit genauso viele Kunden das Warteschlangennetz durch die Senke S_1 wie vom Deaktivierer zum Sperrvektor übergehen, d.h.

$$p_{10}\lambda_1 = p_{13}\lambda_1 .$$

Nach Gleichung (5.4e) ist die Rate $p_{10}\lambda_1$ gleich der Rate λ_{05} . Mit dieser Rate werden dem Warteschlangennetz Kunden der Menge \mathcal{E} durch die Quelle Q_2 zugeführt. Es sind also gleichviele Kunden der Menge \mathcal{E} wie Kunden der Menge \mathcal{T} von der Ausführungseinheit zur Kontrolleinheit übergegangen, d.h. $\lambda_1 = 2\lambda_{05}$. In den Schaltmaschinen werden pro Zeiteinheit die gleiche Anzahl

von Kunden der Menge \mathcal{E} wie der Menge \mathcal{T} bedient. Die PAPER-Architektur bearbeitet genauso viele externe Ereignisse wie Transitionen.

Ist dagegen $t_a = 0$, gehen alle am Deaktivierer ankommenden Kunden mit der Rate λ_1 in die Senke S_1 über ($p_{10} = 1$). Am Deaktivierer kommen nur Kunden der Menge \mathcal{E} mit einer Rate von λ_{05} an. Die PAPER-Architektur bearbeitet nur die externen Ereignisse. Es erfolgt aber keine Reaktion durch das Protokoll, da keine Transitionen aktiviert wurden.

6.2.3 Die Verzweigungswahrscheinlichkeiten der Sperreinheit

Die Verzweigungswahrscheinlichkeiten der Sperreinheit

$$p'_{30} = \frac{\delta_g}{1 + \delta_f} \quad p''_{30} = \frac{\delta_f}{1 + \delta_f} \quad p_{34} = \frac{\delta_f}{1 + \delta_f}$$

werden indirekt von t_a beeinflusst. Sie sind in erster Linie von dem Verhältnis von aktivierten zu schaltfähigen (δ_f), bzw. gesperrten zu schaltfähigen (δ_g) Transitionen abhängig. Diese Verhältnisse wurden im Abschnitt 5.3.1 definiert.

Für die Verzweigungswahrscheinlichkeiten gilt nach den Gleichungen (5.20): $p''_{30} = p_{34}$. Die Wahrscheinlichkeit, daß Kunden von der Sperreinheit zur Senke S_3 übergehen (p''_{30}) ist gleich der Wahrscheinlichkeit, daß Kunden von der Sperreinheit zu ihrer Abarbeitung in die Ausführungseinheit übergehen (p_{34}). Diese Gleichheit liegt in der Konstruktion des Warteschlangennetzes. Die Sperreinheit modelliert sowohl den Sperrvektor als auch den Aktivierer der PAPER-Architektur. Von der Wahrscheinlichkeit p_{34} sind nur die Transitionen betroffen, die aktiviert werden konnten. Diese Transitionen haben nach dem Schaltvorgang einen erneuten Zugriff auf den Sperrvektor um die durch ihr Schalten gesperrten Transitionen zu entsperren. Als deaktivierte Transitionen haben sie keine weitere Funktion mehr. Das wird im Warteschlangennetz durch p''_{30} und die Senke S_3 modelliert.

Sind $\delta_f = \delta_g = 0.5$, dann sind die Verzweigungswahrscheinlichkeiten gleich. In der PAPER-Architektur sind von den durch den Auswerter als schaltfähig erkannten Transitionen eine Hälfte gesperrt, während die andere Hälfte aktiviert wird.

6.2.4 Die Verzweigungswahrscheinlichkeiten der Schaltmaschinen

Im Abschnitt 5.4 wurden die Berechnungsvorschriften für die Verzweigungswahrscheinlichkeiten der Schaltmaschinen hergeleitet.

$$p_{54} \stackrel{(5.22)}{=} \frac{1}{1 + p_{glob}} \quad p_{56} \stackrel{(5.21)}{=} \frac{p_{glob}}{1 + p_{glob}}$$

Diese sind von der Wahrscheinlichkeit p_{glob} abhängig, mit der die Kunden während ihrer Bedienung durch die Schaltmaschinen auf den globalen Speicher zugreifen. Die Abbildung 6.5 zeigt den Verlauf der Verzweigungswahrscheinlichkeiten p_{54} und p_{56} in Bezug auf p_{glob} .

Benutzt *kein* Kunde den globalen Speicher, so ist $p_{56} = 0$. Benutzt dagegen jeder Kunde den globalen Speicher, sind die Verzweigungswahrscheinlichkeiten gleich: $p_{54} = p_{56} = 0.5$. Jeder

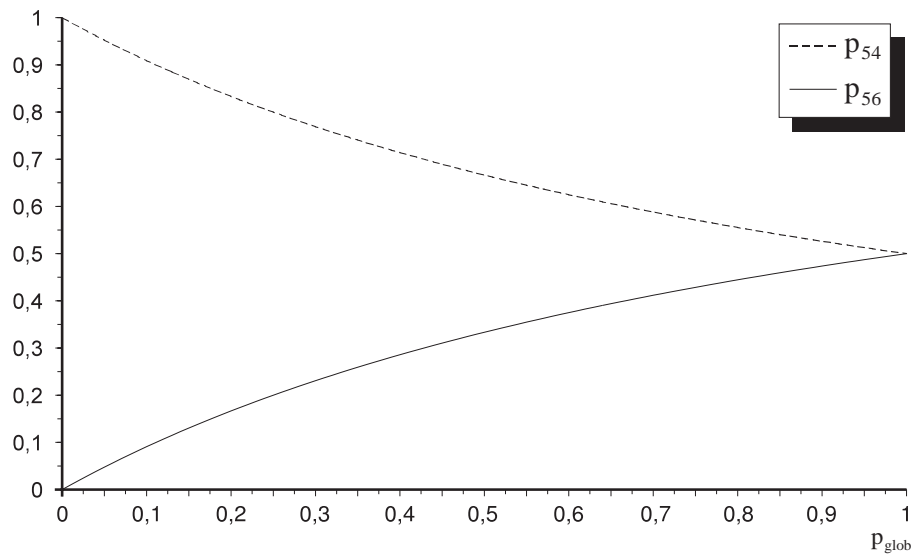


Abbildung 6.5: Der Verlauf der Verzweigungswahrscheinlichkeiten p_{54} und p_{56}

neue Kunde der Mengen \mathcal{T} und \mathcal{E} , der mit der Rate $\lambda_5 \Leftrightarrow \lambda_6 \stackrel{(5.3)}{=} p_{45}\lambda_4 + \lambda_{05}$ an den Schaltmaschinen ankommt, benutzt während seiner Abarbeitung durch die Schaltmaschinen den globalen Speicher. Dadurch werden den Schaltmaschinen vom globalen Speicher die gleiche Anzahl von Kunden zugeführt. Das erklärt die Gleichheit der beiden Verzweigungswahrscheinlichkeiten.

Mit der Veränderung von p_{glob} ändert sich die aggregierte Ankunftsrate λ_5 der Schaltmaschinen. Aus den gerade beschriebenen Gründen führt das bei $p_{glob} = 1$ zu einer Verdopplung der Ankunftsrate. In der Abbildung 6.6 wird der Zusammenhang zwischen einem sich ändernden p_{glob} und der aggregierten Ankunftsrate der Schaltmaschinen dargestellt.

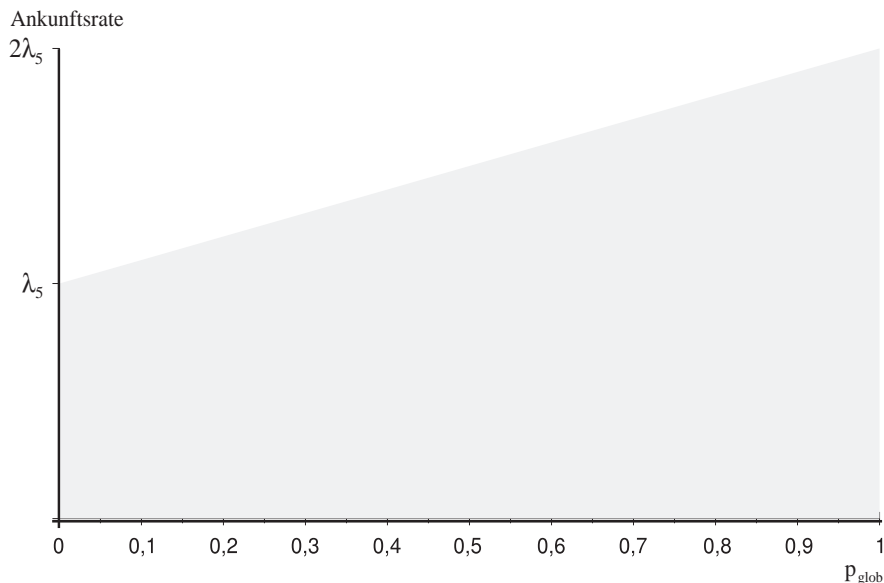


Abbildung 6.6: Das Verhalten von λ_5 bei sich änderndem p_{glob}

Die Bedeutung des globalen Speichers wirkt sich auch auf die Bedienzeit der Schaltmaschinen aus. Diese sind als M/M/c-FCFS PR-Wartesystem modelliert worden. Die Bedienzeit ist somit

exponentiell verteilt mit der Rate μ_5 und alle Kunden haben die gleiche mittlere Bedienzeit. Benutzt ein Kunde nun während seiner Bedienung durch die Schaltmaschinen den globalen Speicher, so wird er im Warteschlangennetz nach dem Speicherzugriff den Schaltmaschinen erneut zugeführt. Geht man von einer konstanten Bedienzeit je Kunden aus, so würde ein Kunde in $2\frac{1}{\mu_5}$ Zeiteinheiten bedient werden. Dieser Wert entspricht aber nicht der Realität in der PAPER-Architektur, da sich durch einen Speicherzugriff die Bearbeitungszeit einer Transition nur um die Dauer des Speicherzugriffes ändert. Aus diesem Grund muß die Bedienrate der Schaltmaschinen im Warteschlangennetz in Abhängigkeit von p_{glob} korrigiert werden.

Vom Markenspeicher und der Quelle Q_2 kommen pro Zeiteinheit im Mittel $p_{45}\lambda_4 + \lambda_{05}$ Kunden an den Schaltmaschinen an. Diese Kunden benutzen mit der Wahrscheinlichkeit p_{glob} den globalen Speicher, was sich durch $p_{56}\lambda_5 = \lambda_6$ ausdrückt. Es werden also $p_{45}\lambda_4 + \lambda_{05} \Leftrightarrow \lambda_6$ Kunden von den Schaltmaschinen bearbeitet, die den globalen Speicher *nicht* benutzen. Diese haben eine mittlere Bedienzeit von $\frac{1}{\mu_5}$. Die verbleibenden λ_6 Kunden und die vom globalen Speicher mit der Rate λ_6 ankommenden Kunden frequentieren die Ausführungseinheiten je einmal. Ihre Bedienzeit darf somit bei jeder Benutzung der Schaltmaschinen nur $\frac{1}{2\mu_5}$ sein. Für die Bedienzeit *aller* λ_5 Kunden, die pro Zeiteinheit an den Schaltmaschinen ankommen, gilt somit:

$$[(p_{45}\lambda_4 + \lambda_{05}) \Leftrightarrow \lambda_6] \frac{1}{\mu_5} + 2\lambda_6 \frac{1}{2\mu_5} \stackrel{(5.4d)}{=} (\lambda_1 \Leftrightarrow \lambda_6) \frac{1}{\mu_5} + \frac{\lambda_6}{\mu_5} = \frac{\lambda_1}{\mu_5} .$$

Jeder einzelne Kunde hat dann eine gemittelte Bedienzeit von

$$\frac{1}{\bar{\mu}_5} = \frac{\lambda_1}{\lambda_5\mu_5} \stackrel{(5.15)}{=} \frac{p_{54}}{\mu_5} .$$

Das entspricht einer Bedienrate von

$$\bar{\mu}_5 = \frac{\mu_5}{p_{54}} \stackrel{(5.22)}{=} (1 + p_{glob})\mu_5 \quad (6.3)$$

Somit ist $(1 + p_{glob})$ der Korrekturfaktor für die Bedienrate des M/M/c-FCFS PR-Wartesystems der Schaltmaschinen wenn die Kunden mit der Wahrscheinlichkeit p_{glob} während ihrer Bearbeitung auf den globalen Speicher zugreifen.

6.3 Die Ankunftsraten der Prioritätsklassen

Die Schaltmaschinen wurden als M/M/c-FCFS PR-Wartesystem mit zwei Prioritätsklassen $\mathcal{P} = \{1, 2\}$ modelliert. Die Kunden der Menge \mathcal{E} , die mit der Rate λ_{05} an den Schaltmaschinen ankommen, bilden die erste Prioritätsklasse. Sie haben Vorrang vor allen anderen Kunden, die mit einer Rate von $\lambda_5 \Leftrightarrow \lambda_{05}$ an den Schaltmaschinen ankommen. Für die Ankunftsraten der beiden Prioritätsklassen gilt:

$$\begin{aligned} \lambda_5^{P1} &= \lambda_{05} , \\ \lambda_5^{P2} &\stackrel{(5.23)}{=} \frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \cdot (1 + p_{glob}) \cdot \lambda_{05} \Leftrightarrow \lambda_{05} \stackrel{(5.18)}{=} \frac{p_{glob} + t_a}{1 \Leftrightarrow t_a} \cdot \lambda_{05} . \end{aligned} \quad (6.4)$$

Im folgenden soll das Verhalten der Ankunftsraten λ_5^{P1} und λ_5^{P2} zueinander untersucht werden. Die

hier gemachten Betrachtungen über das Verhältnis der Ankunftsrate der beiden Prioritätsklassen sind eher theoretischer Natur. Der Kreislauf von Kunden, die im Warteschlangennetz auf den globalen Speicher zugreifen, ist in der PAPER-Architektur nicht existent. Er dient zur Modellierung der Verzögerung bei der Abarbeitung von Transitionen oder externen Ereignissen, die durch einen Zugriff auf den globalen Speicher entstehen (\diamond Abbildung 3.2).

Sei ω das Verhältnis der Ankunftsrate der Prioritätsklasse 1 zur Ankunftsrate der Prioritätsklasse 2:

$$\omega \stackrel{\text{def.}}{=} \frac{\lambda_5^{P1}}{\lambda_5^{P2}} \stackrel{(6.4)}{=} \frac{1 \Leftrightarrow t_a}{p_{glob} + t_a} \quad (6.5)$$

Nun können zwei Fälle unterschieden werden:

1. Kein Kunde benutzt den globalen Speicher ($p_{glob} = 0$, $\omega = \frac{1-t_a}{t_a}$)

Werden keine Transitionen aktiviert, so ist $t_a = 0$. In diesem Fall bearbeiten die Schaltmaschinen nur Kunden der Prioritätsklasse 1, also externe Ereignisse. Es werden keine Protokolltransitionen bearbeitet. Den Verlauf von ω in Bezug auf $t_a \in (0, 1)$ zeigt Abbildung 6.7.

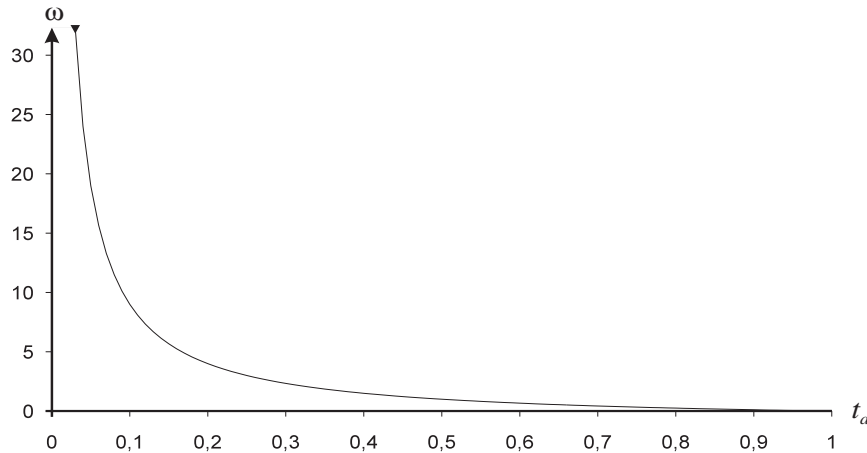


Abbildung 6.7: Das Verhältnis ω für $p_{glob} = 0$ und $t_a \in (0, 1)$

Das Verhältnis von Kunden der Prioritätsklasse 1 und denen der Prioritätsklasse 2 ist gleich, wenn $\omega = 1$ ist. In diesem Fall ist $t_a = \frac{1}{2}$. Für $t_a < \frac{1}{2}$ wird die von den Schaltmaschinen zu bearbeitende Kundenmenge von Kunden der Prioritätsklasse 1 dominiert. Die Schaltmaschinen bearbeiten hauptsächlich Kunden der Prioritätsklasse 2 wenn $t_a > \frac{1}{2}$. Nähert sich t_a dem Wert 1, so ist der Anteil von Kunden der Prioritätsklasse 1 an der zu bearbeitenden Menge von Kunden verschwindend gering.

2. Die Kunden benutzen den globalen Speicher ($p_{glob} \neq 0$)

Wenn keine Transitionen aktiviert werden ($t_a = 0$), gilt für ω : $\omega = \frac{1}{p_{glob}}$. Vom Markenspeicher gehen keine Kunden der Menge \mathcal{T} (Prioritätsklasse 2) zu den Schaltmaschinen über ($p_{45} = 0$). Die Kunden der Menge \mathcal{E} (Prioritätsklasse 1) benutzen mit der Wahrscheinlichkeit p_{glob} den globalen Speicher. Nach der Konstruktion des Warteschlangennetzes werden Kunden

der Menge \mathcal{E} nur bei ihrem Auftreten bevorzugt behandelt. Greifen sie während ihrer Bearbeitung auf den globalen Speicher zu, so haben sie bei ihrer erneuten Ankunft an den Schaltmaschinen ihre Prioritätsklasse geändert. Aus diesem Grund ist im Falle $t_a = 0$ die Ankunftsrate von Kunden der Prioritätsklasse 1 das $\frac{1}{p_{glob}}$ -fache der Ankunftsrate von Kunden der Prioritätsklasse 2. In der Abbildung 6.8 wird der Verlauf von ω für verschiedene p_{glob} dargestellt.

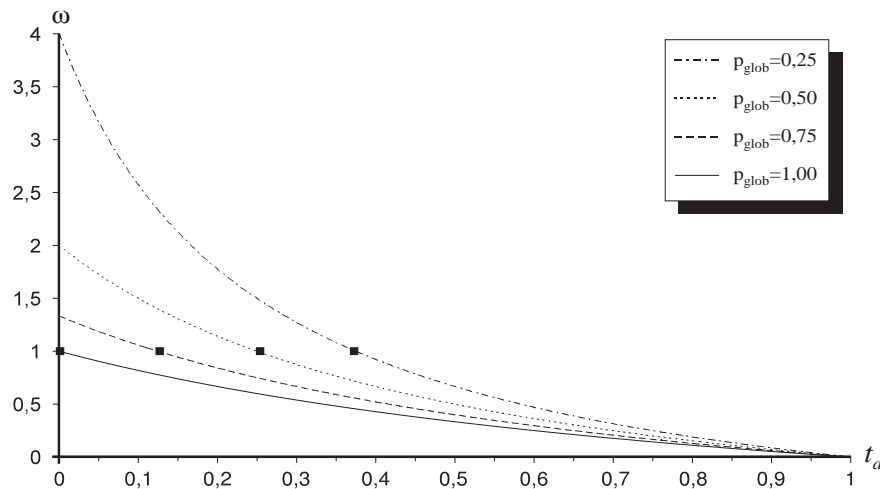


Abbildung 6.8: Das Verhältnis ω für verschiedene p_{glob} und $t_a \in [0, 1)$

Eine Gleichheit für Kunden der Prioritätsklasse 1 und der Prioritätsklasse 2 tritt für $t_a = \frac{1-p_{glob}}{2}$ ein. Bei Benutzung des globalen Speichers tritt die Gleichheit für die Ankunftsraten der beiden Prioritätsklassen je nach Größe von p_{glob} für $t_a < \frac{1}{2}$ ein. Je größer der Wert von p_{glob} ist, umso kleiner ist dann der Wert von t_a für die Gleichheit der Ankunftsraten. Die Kunden der Prioritätsklasse 2 dominieren die Kunden der Prioritätsklasse 1 schon für $t_a < \frac{1}{2}$ (Im Falle $p_{glob} = 0$ erst für $t_a > \frac{1}{2}$).

Durch ω kann das Verhältnis der Ankunftsraten für die beiden Prioritätsklassen ermittelt werden. Dadurch lassen sich in der PAPER-Architektur die Verzögerungen im Schaltvorgang einer Transition abschätzen, die durch die bevorzugte Behandlung der externen Ereignisse entstehen. So läßt sich beispielsweise durch die Prognose von ω die Anzahl der Schaltmaschinen so abstimmen, daß durch die bevorzugte Bearbeitung der externen Ereignisse keine Verzögerung für den Schaltvorgang der Transitionen entsteht.

7. Die Bewertung der PAPER-Architektur durch das Warteschlangenmodell

In den vorangegangenen Kapiteln wurden die Grundlagen für eine quantitative Bewertung der PAPER-Architektur auf der Basis eines Jackson-Warteschlangennetzes geschaffen. Für eine quantitative Bewertung werden die Bearbeitungszeiten der einzelnen Komponenten in der PAPER-Architektur unter realen Bedingungen bei der Ausführung eines Kommunikationsprotokolls benötigt.

Für die PAPER-Architektur wurde im Rahmen des Forschungsprojekts PIKOM ein Emulator entwickelt. Auf diesem erfolgt zur Zeit eine Implementierung des Kommunikationsprotokolls XTP¹ [XTP92], welches für den Einsatz in heutigen Hochgeschwindigkeitsnetzen entwickelt wurde. XTP umfaßt die dritte und vierte Ebene des ISO-Referenzmodells. Das ist zum einen die Netzebene (Network Layer), die für den Verbindungsauf- und Verbindungsabbau zuständig ist, und die Transportebene (Transport Layer), die als netzunabhängige *End-to-End*-Verbindung den Datenfluß gewährleisten soll. XTP wird durch erweiterte endliche Automaten beschrieben, die die unterschiedlichen Protokollaufgaben realisieren. Von diesen sogenannten *Statusmaschinen* wurden bis zur Fertigstellung dieser Arbeit sechs² in ein PENCIL-Netz übertragen, welches als PENCIL/C-Programm auf dem PAPER-Emulator zur Ausführung gebracht wurde. Das Senden und Empfangen von Datenpaketen wurde dabei durch zusätzliche Transitionen simuliert. Die bei der Emulation auf einem Sparc 10 Rechner ermittelten Werte dienen als Basis für eine quantitative Bewertung mittels des im Kapitel 5 konstruierten Warteschlangennetzes.

7.1 Die benötigten Basisgrößen für die Bewertung

Zur quantitativen Analyse der PAPER-Architektur durch das im Kapitel 5 konstruierte Warteschlangennetz sind zunächst die Ankunftsraten für die einzelnen Knoten zu ermitteln. Durch diese Raten werden die Poisson'schen Ankunftsprozesse und damit die auftretende Last an den Knoten des Warteschlangennetzes modelliert. Zur Ermittlung der Ankunftsraten und Verzweigungswahrscheinlichkeiten durch die im Abschnitt 5.3 hergeleiteten Berechnungsvorschriften sind Informationen über die Struktur des PENCIL-Netzes notwendig.

Die Emulatorläufe für XTP haben gezeigt, daß die Bearbeitung einer Transition oder eines externen Ereignisses zur Folge hat, daß der Auswerter im Mittel 21,5 potentiell schaltfähige

¹ eXpress Transport Protocol

² Context-Manager, Input-Manager, Output-Manager, Control-Send-Machine, Read-Close-Machine, Write-Close-Machine

Transitionen auf ihre Schaltfähigkeit prüfen muß. Diese hohe Anzahl läßt sich durch die bei der Übertragung der erweiterten endlichen Automaten in ein PENCIL-Netz entstandene hohe Konnektivität des Netzes erklären. An einer Reduzierung der Konnektivität des PENCIL-Netzes für XTP wird im Rahmen des Forschungsprojekts PIKOM augenblicklich gearbeitet. Von den potentiell schaltfähigen Transitionen sind im Mittel 5,8% auch schaltfähig und durch den Aktivierer auf ihre Aktivierbarkeit zu prüfen. Dabei stellte sich heraus, daß von den schaltfähigen Transitionen im Mittel 77% aktiviert werden. Übertragen auf die im Abschnitt 5.3 eingeführten Schreibweisen ergeben sich die in der Tabelle 7.1 aufgeführten Werte.

γ	21,5
p_{23}	0,058
t_s	1,247
δ_f	0,77
δ_g	0,23
t_a	0,96019

Tabelle 7.1: Die vom PENCIL-Netz abhängigen Werte zur Berechnung der Ankunfts-raten und Verzweigungswahrscheinlichkeiten

Die Bearbeitung *eines* Deaktivierungspakets durch den Deaktivierer hat zur Folge, daß im Mittel $t_a = 0,96019$ Aktivierungspakete entstehen. Von den potentiell schaltfähigen Transitionen werden im Mittel also nur 4,4% aktiviert. Dieser relativ kleine Prozentsatz ist täuschend, da durch die hohe Konnektivität viele Transitionen potentiell schaltfähig werden. Die Tatsache, daß im Mittel zu fast jeder Transition oder externem Ereignis eine Transition aktiviert wird, ist ein wünschenswerter Zustand. Dadurch wird eine zügige Protokollverarbeitung gewährleistet.

Mit dem Wert $t_a = 0,96019$ sind Aussagen über das Verhalten des Warteschlangennetzes möglich. Wie im Abschnitt 6.2 ausgeführt wurde, sind die Verzweigungswahrscheinlichkeiten des Markenspeichers p_{41} und p_{45} für $t_a \rightarrow 1$ nahezu identisch. Das bedeutet, daß nach den dortigen Ausführungen von der PAPER-Architektur hauptsächlich Transitionen des PENCIL-Netzes bearbeitet werden. Die Anzahl der externen Ereignisse ist im Vergleich zu den im Umlauf befindlichen Transitionen gering. Dies spiegelt sich im Verhältnis der Ankunfts-raten für die beiden Prioritäts-klassen an den Schaltmaschinen wieder. Für $t_a = 0,96019$ ist das in Gleichung (6.5) definierte Verhältnis ω der Ankunfts-raten λ_5^{P1} und λ_5^{P2} für die Prioritätsklassen an den Schaltmaschinen in Tabelle 7.2 aufgeführt.

	ω
$p_{glob} = 0$	0,04
$p_{glob} = 1$	0,02

Tabelle 7.2: Das Verhältnis der Ankunfts-raten für die beiden Prioritätsklassen an den Schaltmaschinen

In der aggregierten Ankunftsrate λ_5 der Schaltmaschinen sind nur 2% – 4% Kunden der Prioritäts-klasse 1 enthalten. Diese Aussagen bedeuten hinsichtlich der betrachteten XTP-Implementierung,

daß die Bearbeitung von Transitionen durch die bevorzugte Behandlung von externen Ereignissen bei deren Auftreten an den Schaltmaschinen nur unwesentlich verzögert wird.

Mit den Werten aus der Tabelle 7.1 lassen sich nun die **Ankunftsrate**n für die Knoten des Warteschlangennetzes in Abhängigkeit von der externen Ankunftsrate λ_{05} berechnen. Die Benutzung des globalen Speichers wird durch die Extremfälle $p_{glob} = 0$ (es wird kein globaler Speicher benutzt) und $p_{glob} = 1$ (alle Kunden der Schaltmaschinen benutzen den globalen Speicher) berücksichtigt. Eine Betrachtung weiterer Werte von p_{glob} bei der Berechnung der Basisgrößen ist nicht sinnvoll, da durch die Betrachtung dieser Extremfälle die dadurch beeinflussten Größen tendenziell beschrieben werden.

Deaktivierer		λ_1	$25,119\lambda_{05}$
Auswerter		λ_2	$540,065\lambda_{05}$
Sperreinheit		λ_3	$55,443\lambda_{05}$
Markenspeicher		λ_4	$49,237\lambda_{05}$
Schaltmaschinen (aggregiert)	$p_{glob} = 0$	λ_5	$25,119\lambda_{05}$
	$p_{glob} = 1$		$50,238\lambda_{05}$
Schaltmaschinen (Prioritätsklasse 1)		λ_5^{P1}	λ_{05}
Schaltmaschinen (Prioritätsklasse 2)	$p_{glob} = 0$	λ_5^{P2}	$24,119\lambda_{05}$
	$p_{glob} = 1$		$49,238\lambda_{05}$
globaler Speicher	$p_{glob} = 0$	λ_6	$0,000\lambda_{05}$
	$p_{glob} = 1$		$25,119\lambda_{05}$

Tabelle 7.3: Die *Ankunftsrate*n für die Knoten des Warteschlangennetzes

Die Abbildung 7.1 verdeutlicht das Verhältnis der Ankunftsrate n zueinander. Auffallend ist die hohe Ankunftsrate λ_2 des Auswerter s. Das kommt daher, daß λ_2 nach der Konstruktion des Warteschlangennetzes das γ -fache der Ankunftsrate λ_1 des Deaktivierers ist. Im vorliegenden Fall der XTP-Implementierung hat das PENCIL-Netz einen hohen Konnektivitätsgrad, der sich in $\gamma = 21,5$ widerspiegelt.

Neben den Ankunftsrate n lassen sich mit den Werten aus der Tabelle 7.1 auch die **Verzweigungswahrscheinlichkeiten** des Warteschlangennetzes berechnen, die in Tabelle 7.4 aufgeführt sind.

Aus den Emulatorläufen sind die Zeiten bekannt, die bei der Abarbeitung von XTP durch die PAPER-Architektur in den einzelnen Komponenten durchschnittlich benötigt werden (\diamond Tabelle 7.5). Diese Bearbeitungszeiten entsprechen selbstverständlich nicht den Werten, wie sie bei einer Hardware-Implementierung von PAPER auftreten würden. Sie geben aber dennoch die Möglichkeit, das Verhalten des Warteschlangennetzes zu analysieren und damit Rückschlüsse für die PAPER-Architektur zu ziehen. Der Zugriff auf den globalen Speicher soll zunächst nicht berücksichtigt werden. Er wird später Gegenstand einer gesonderten Betrachtung sein.

Mit diesen Bearbeitungszeiten lassen sich die **Bedienrate**n für die Knoten des Warteschlangennetzes ermitteln. Das Warteschlangennetz wurde so konstruiert, daß die Bedienzeiten aller Knoten exponentiell verteilt sind. Die angegebenen Bearbeitungszeiten können als die mittlere

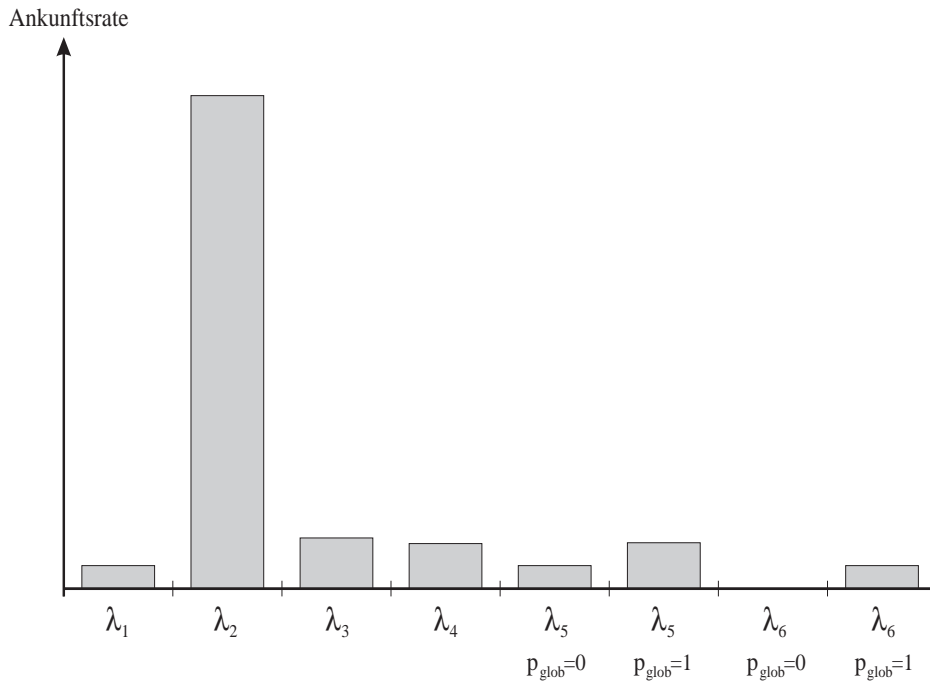


Abbildung 7.1: Das größenmäßige Verhältnis der *Ankunftsrate*n zueinander

Deaktivierer		Auswerter		Sperrereinheit		Markenspeicher	
p_{10}	0,040	p_{20}	0,942	p'_{30}	0,130	p_{41}	0,510
p_{13}	0,960	p_{23}	0,058	p''_{30}	0,435	p_{45}	0,490
				p_{34}	0,435		

Schaltmaschinen				globaler Speicher	
$p_{glob} = 0$		$p_{glob} = 1$			
p_{54}	1,0	p_{54}	0,5	p_{65}	1,0
p_{56}	0,0	p_{56}	0,5		

Tabelle 7.4: Die *Verzweigungswahrscheinlichkeiten* im Warteschlangennetz

ren Bedienzeiten der Knoten im Warteschlangennetz aufgefaßt werden. Für die entsprechenden Bedienraten gilt nach Gleichung (4.26d): $\mu_i = \frac{1}{s_i}$. Die Bedienrate für die Sperrereinheit und den Markenspeicher sind dabei *gesondert* zu behandeln.

Die **Sperrereinheit** erfüllt im Warteschlangennetz *zwei* Funktionen. Zum einen modelliert sie den Aktivierungsvorgang einer als schaltfähig erkannten Transition und zum anderen die zu synchronisierenden Zugriffe von Aktivierer und Deaktivierer auf den Sperrvektor. Die Bedienrate der Sperrereinheit muß diesem Sachverhalt gerecht werden. Da es sich um einen exponentiell verteilten Bedienprozeß handelt, müssen die Bedienzeiten für die verschiedenen Funktionen der Sperrereinheit *gleich* sein. Für die Aktivierung einer Transition – inklusive dem Zugriff auf den Sperrvektor – ist die gesamte Bedienzeit $s_3 + s'_3$. Diese Bedienzeit wird von $p_{23}\lambda_2$ Kunden pro Zeiteinheit in Anspruch genommen. Für das Entsperren von Transitionen, die durch die Aktivierung einer anderen Transition gesperrt wurden, werden s'_1 Zeiteinheiten von $p_{13}\lambda_1$ Kunden

		Zeit in μs
Deaktivierer	s_1	280,64
Zugriff des Deaktivierers auf den Sperrvektor	s'_1	13,12
Auswerter	s_2	17,89
Aktivierer	s_3	166,27
Zugriff des Aktivierers auf den Sperrvektor	s'_3	8,54
Markenspeicherzugriff während der Aktivierungsphase	s'_4	37,28
Markenspeicherzugriff während der Deaktivierungsphase	s''_4	75,09
Schaltmaschinen	s_5	516,20

Tabelle 7.5: Die durchschnittlichen *Bearbeitungszeiten* der einzelnen Komponenten bei der Emulation der PAPER-Architektur

benötigt. Somit kann die Bedienzeit der Sperreinheit durch

$$\bar{s}_3 = \frac{p_{23}\lambda_2 \cdot (s_3 + s'_3) + p_{13}\lambda_1 \cdot s'_1}{p_{23}\lambda_2 + p_{13}\lambda_1} \stackrel{(5.3)}{=} \frac{p_{23}\lambda_2 \cdot (s_3 + s'_3) + p_{13}\lambda_1 \cdot s'_1}{\lambda_3} \stackrel{(B)}{=} \frac{s_3 + s'_3 + \delta_f \cdot s'_1}{1 + \delta_f} \quad (7.1a)$$

ermittelt werden. Für die Bedienrate der Sperreinheit gilt dann

$$\mu_3 = \frac{1}{\bar{s}_3} \stackrel{(7.1a)}{=} \frac{1 + \delta_f}{s_3 + s'_3 + \delta_f \cdot s'_1} \quad (7.1b)$$

Bei der Ermittlung der Bedienrate für den **Markenspeicher** wird analog vorgegangen. Dieser modelliert sowohl die Aktivierungs- als auch die Deaktivierungsphase im Schaltvorgang einer Transition. Für die mittlere Bedienzeit gilt dann

$$\bar{s}_4 = \frac{\overbrace{p_{34}\lambda_3 \cdot s'_4}^{\text{Aktivierungsphase}} + \overbrace{p_{54}\lambda_5 \cdot s''_4}^{\text{Deaktivierungsphase}}}{p_{34}\lambda_3 + p_{54}\lambda_5} \stackrel{(5.3)}{=} \frac{p_{34}\lambda_3 \cdot s'_4 + p_{54}\lambda_5 \cdot s''_4}{\lambda_4} \stackrel{(B)}{=} \frac{t_a \cdot s'_4 + s''_4}{t_a + 1} \quad (7.2a)$$

Die Bedienrate des Markenspeichers läßt sich dann durch Kehrwertbildung aus Gleichung (7.2a) ermitteln.

$$\mu_4 = \frac{1}{\bar{s}_4} \stackrel{(7.2a)}{=} \frac{t_a + 1}{t_a \cdot s'_4 + s''_4} \quad (7.2b)$$

Die Tabelle 7.6 zeigt die ermittelten Bedienraten des Warteschlangennetzes auf der Basis von Millisekunden.

Damit das Warteschlangennetz eine Produktformlösung hat, muß es die Local-Balance-Eigenschaft besitzen. Diese ist gegeben, wenn sämtliche Knoten stationäre Zustände einnehmen können: $\rho_i \stackrel{(4.40)}{=} \frac{\lambda_i}{c_i \cdot \mu_i} < 1, \forall i \in \mathcal{N}$.

Alle Ankunftsrate sind von der externen Ankunftsrate λ_{05} abhängig (\diamond Tabelle B.1). Daher sind auch die Auslastungen ρ_i der einzelnen Knoten von λ_{05} abhängig. Ob die einzelnen Knoten stationäre Zustände einnehmen können, wird somit von λ_{05} bestimmt. Die Tabelle 7.7 gibt die

		Rate pro <i>ms</i>
Deaktivierer	μ_1	3,563
Auswerter	μ_2	55,897
Sperreinheit	μ_3	9,572
Markenspeicher	μ_4	17,678
Schaltmaschinen	$p_{glob} = 0$	1,937
	$p_{glob} = 1$	3,874

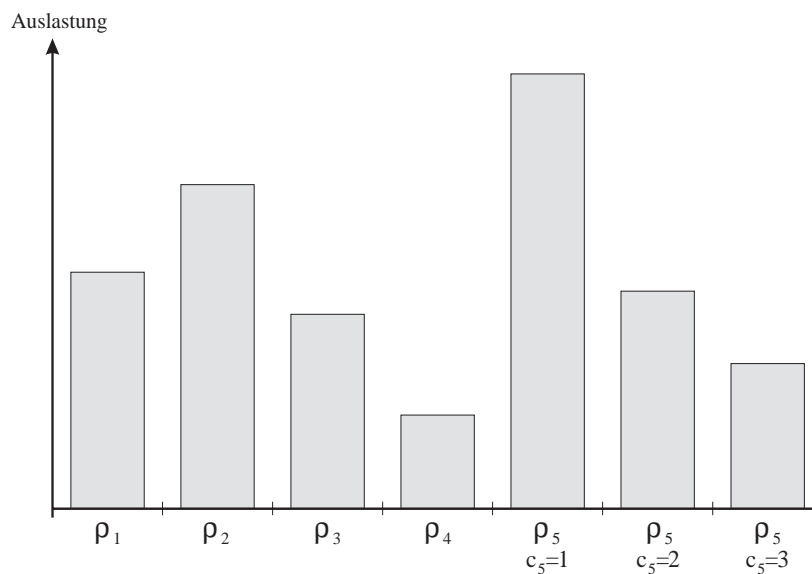
Tabelle 7.6: Die *Bedienraten* des Warteschlangennetzes

Auslastungen der einzelnen Knoten im Warteschlangennetz der PAPER-Architektur an.

		Typ	Rate pro <i>ms</i>		
Deaktivierer	M/M/1	ρ_1	$7,049\lambda_{05}$	$\lambda_{05} <$	0,142
Auswerter	M/M/1	ρ_2	$9,662\lambda_{05}$	$\lambda_{05} <$	0,104
Sperreinheit	M/M/1	ρ_3	$5,792\lambda_{05}$	$\lambda_{05} <$	0,173
Markenspeicher	M/M/1	ρ_4	$2,785\lambda_{05}$	$\lambda_{05} <$	0,359
Schaltmaschinen	$p_{glob} = 0$	M/M/ c_5	$\frac{1}{c_5} 12,967\lambda_{05}$	$\lambda_{05} < c_5 \cdot$	0,077
	$p_{glob} = 1$		$\frac{1}{c_5} 12,967\lambda_{05}$	$\lambda_{05} < c_5 \cdot$	0,077

Tabelle 7.7: Die *Auslastungen* ρ_i der Knoten des Warteschlangennetzes und die Bedingungen für die Existenz stationärer Zustände

Um die Leistungsgrößen für die Knoten des Warteschlangennetzes berechnen zu können, muß dieses die Local-Balance-Eigenschaft besitzen. Aus diesem Grunde muß für die externe Ankunftsrate $\lambda_{05} < \min_{i \in \mathcal{N}} \left\{ \frac{1}{\rho_i} \right\}$ gelten.

Abbildung 7.2: Das größenmäßige Verhältnis der *Auslastungen* zueinander

Die Abbildung 7.2 zeigt die Größenverhältnisse der Auslastungen untereinander. Da die Schaltmaschinen als M/M/ c -Wartesystem modelliert wurden, ist ihre Auslastung beispielhaft für den

Einsatz mehrerer Bedieneinheiten dargestellt. In der Kontrolleinheit hat der Auswerter die höchste Auslastung. Dies wird durch die hohe Anzahl der indirekt inzidenten Transitionen im PENCIL-Netz von XTP hervorgerufen.

Auf der Basis der hier ermittelten Werte erfolgen die weiteren Analysen im Verlaufe dieses Kapitels.

7.2 Die Dimensionierung der Verbindungswarteschlangen

Die Kontroll- und die Ausführungseinheit der PAPER-Architektur sind durch zwei FIFO-Warteschlangen miteinander verbunden. Sie dienen zur Aufnahme von Aktivierungs- und Deaktivierungspaketen bei einer zeitlichen Verzögerung im Zusammenwirken von Kontroll- und Ausführungseinheit. Damit eine reibungslose Protokollverarbeitung gewährleistet ist, muß die Anzahl der Warteplätze an das zu erwartende Lastaufkommen in der PAPER-Architektur angepaßt werden.

Die Verbindungswarteschlangen werden im Warteschlangennetz der PAPER-Architektur durch die Warteschlange des Deaktivierers für die Richtung *Ausführungseinheit*→*Kontrolleinheit* und die Warteschlange des Markenspeichers für die Richtung *Kontrolleinheit*→*Ausführungseinheit* modelliert. Letztere wird sowohl von Kunden der Aktivierungs- als auch der Deaktivierungsphase benutzt. Für eine Größenanpassung im Hinblick auf die PAPER-Architektur sind nur die Kunden der Aktivierungsphase interessant.

Im Warteschlangennetz der PAPER-Architektur sind sowohl der Deaktivierer als auch der Markenspeicher als M/M/1-Wartesystem modelliert. Die Anzahl der Warteplätze läßt sich durch die Wahrscheinlichkeit, daß sich mindestens k Kunden in dem M/M/1-Wartesystem befinden berechnen. Nach Gleichung (4.26h) gilt dafür: $P\{n \geq k\} = \rho^k$. ρ^k ist die Wahrscheinlichkeit, daß ein ankommender Kunde abgewiesen werden muß, wenn sich schon k Kunden im M/M/1-Wartesystem befinden. Das Ziel ist nun, diese Wahrscheinlichkeit möglichst klein zu halten, d.h.

$$\rho^k \leq \varepsilon, \quad \varepsilon \in [0, 1].$$

Die benötigte Anzahl von Warteschlangenplätzen ergibt sich dann durch

$$\rho^k \leq \varepsilon \quad \Leftrightarrow \quad k > \frac{\log \varepsilon}{\log \rho}, \quad \rho \in (0, 1). \quad (7.3)$$

Geht man nun davon aus, daß $\varepsilon = 10^{-x}$, $x \geq 0$, ist, dann ist

$$\log \varepsilon = \log 10^{-x} = \Leftrightarrow x \log 10, \quad x \geq 0.$$

Unter Benutzung des dekadischen Logarithmus gilt für Gleichung (7.3):

$$k > \frac{\Leftrightarrow x}{\log_{10} \rho}, \quad x \geq 0, \quad \rho \in (0, 1). \quad (7.4)$$

k nimmt immer positive Werte an, da für $\rho \in (0, 1)$ $\log_{10} \rho < 0$ ist. Der Wert von k wächst bei konstanter Auslastung ρ linear mit dem Exponenten der Zehnerpotenz von ε . Die Abbildung 7.3 zeigt das Verhalten von k als Funktion von ρ für verschiedene ε .

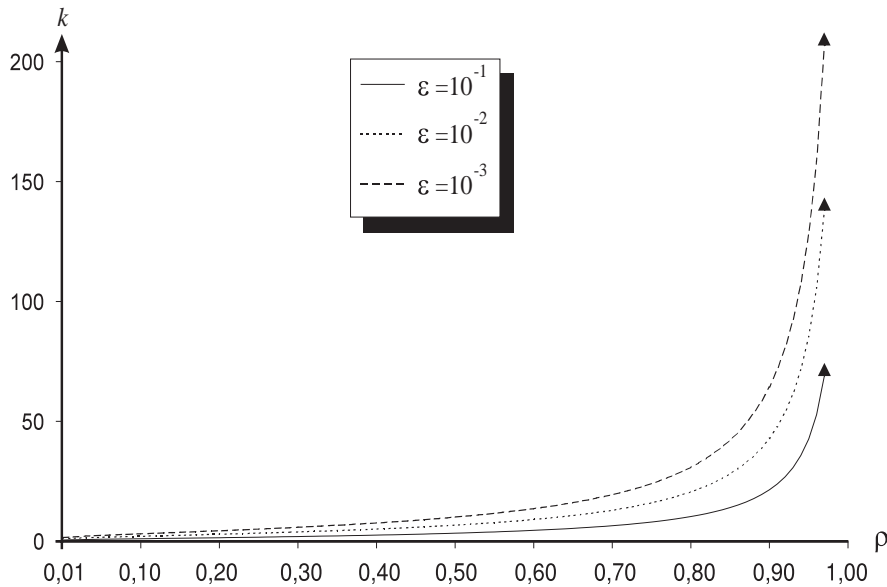


Abbildung 7.3: Die Anzahl der benötigten Warteplätze in Abhängigkeit der Auslastung

Die Abbildung läßt erkennen, daß bis zu einer Auslastung von 50% ($\rho = 0,5$) nur eine geringe Anzahl von Warteplätzen benötigt wird. Dieses Verhalten spiegelt sich auch in der mittleren Warteschlangenlänge $E[n_q]$ eines M/M/1-Wartesystems wieder (\Leftrightarrow Abbildung 7.4).

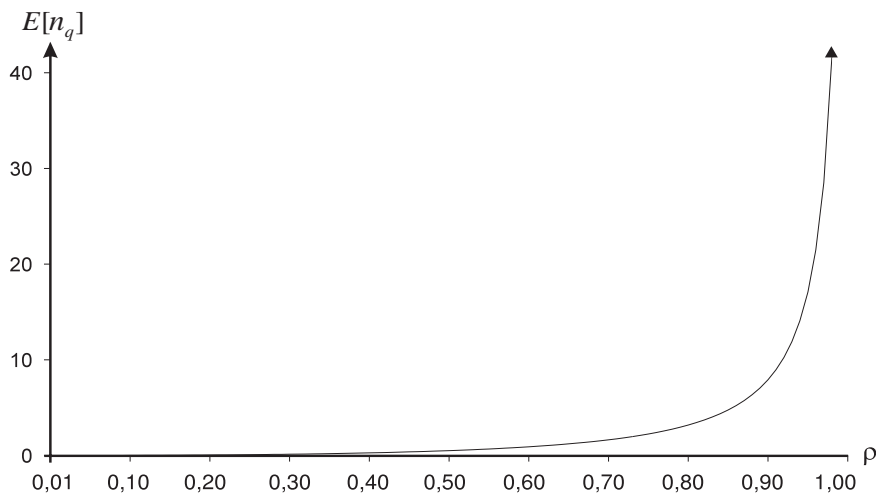


Abbildung 7.4: Die mittlere Warteschlangenlänge eines Markov'schen Wartesystems

Die Verbindungswarteschlange *Ausführungseinheit* \rightarrow *Kontrolleinheit* wird im Warteschlangennetz der PAPER-Architektur durch die Warteschlange des Deaktivierers modelliert. Nach den durch den PAPER-Emulator ermittelten Werte muß für alle Knoten des Warteschlangennetzes $\lambda_{05} < 0,104$ gelten (\Leftrightarrow Tabelle 7.7). Der Deaktivierer ist dann zu maximal 72% ausgelastet. Die Tabelle 7.8 zeigt die benötigte Anzahl von Warteplätzen der Verbindungswarteschlange für die Deaktivierungspakete. ε gibt die Wahrscheinlichkeit an, daß nur jeder 10^x -te Kunde abgewiesen werden muß.

In der Verbindungswarteschlange befinden sich bei dieser Auslastung im Mittel $E_1[n_q] \stackrel{(4.26f)}{=} \dots$

ε		Warteplätze
10^{-1}	$k > 7,194$	8
10^{-2}	$k > 14,388$	15
10^{-3}	$k > 21,582$	22
10^{-4}	$k > 28,776$	29
10^{-5}	$k > 35,960$	36
10^{-6}	$k > 43,164$	44

Tabelle 7.8: Die benötigte Anzahl von Warteplätzen für den *Deaktivierer*

1,925 Deaktivierungspakete, die im Mittel $E_1[q] \stackrel{(4.26e)}{=} 0,744$ Millisekunden auf ihre Bearbeitung warten müssen.

Die Verbindungswarteschlange für die Richtung *Kontrolleinheit* \rightarrow *Ausführungseinheit* wird durch einen Teil der Warteschlange des Markenspeichers modelliert. Für eine Größenanpassung sind nur die Kunden interessant, die sich in der Aktivierungsphase befinden. Diese kommen mit einer Rate von $p_{34}\lambda_3$ am Markenspeicher an. Wegen der Gesetzmäßigkeiten im Kundenfluß des Warteschlangennetzes gilt: $p_{34}\lambda_3 \stackrel{(5.1c)}{=} p_{45}\lambda_4$.

Da im betrachteten Fall der XTP-Implementierung $t_a = 0,96019$ ist, sind die Verzweigungswahrscheinlichkeiten p_{45} und p_{41} annähernd gleich. Die Warteschlange des Markenspeichers setzt sich also zu fast gleichen Teilen aus Kunden der Aktivierungs- und Deaktivierungsphase zusammen. Die Verbindungswarteschlange benötigt somit die *Hälfte* der Warteplätze, wie sie durch Gleichung (7.4) für das M/M/1-Wartesystem des Markenspeichers ermittelt werden können.

Bei der maximal möglichen Ankunftsrate der externen Ereignisse ist der Markenspeicher nur zu 28% ausgelastet. Gemäß der Abbildung 7.3 wird für den Markenspeicher eine geringe Anzahl von Warteplätzen benötigt. Die Tabelle 7.9 zeigt die benötigte Anzahl von Warteplätzen der Verbindungswarteschlange für Aktivierungspakete.

ε		Warteplätze
10^{-1}	$k > 1,801$	1
10^{-2}	$k > 3,603$	2
10^{-3}	$k > 5,404$	3
10^{-4}	$k > 7,206$	4
10^{-5}	$k > 9,007$	5
10^{-6}	$k > 10,809$	6

Tabelle 7.9: Die benötigte Anzahl von Warteplätzen für den *Markenspeicher*

In der Warteschlange des Markenspeichers befinden sich bei dieser Auslastung im Mittel $E_4[n_q] \stackrel{(4.26f)}{=} 0,108$ Kunden. In der Verbindungswarteschlange befinden sich somit im Mittel 0,054 Aktivierungspakete.

Die Verbindungswarteschlange für die Aktivierungspakete wird in der vorliegenden XTP-Implementierung von $p_{34}\lambda_3 \stackrel{(5.1c)}{=} p_{45}\lambda_4 = 24,119\lambda_{05}$ Aktivierungspaketen pro Zeiteinheit be-

nutzt. Im gleichen Zeitraum durchlaufen $\lambda_1 \stackrel{(5.3)}{=} p_{41} \lambda_4 = 25,11932 \lambda_{05}$ Deaktivierungspakete die zugehörige Verbindungswarteschlange. Diese annähernde Gleichheit hat ihre Ursache in der im Abschnitt 6.2.1 beschriebenen Gleichheit von p_{41} und p_{45} für $t_a \rightarrow 1$. Obwohl pro Zeiteinheit fast die gleiche Anzahl von Aktivierungs- bzw. Deaktivierungspaketen die Verbindungswarteschlangen durchlaufen, müssen diese doch in Abhängigkeit der zu erwartenden Last unterschiedlich dimensioniert werden. So werden für die Verbindungswarteschlange der Deaktivierungspakete deutlich mehr Warteplätze benötigt. Die Ursache liegt in den unterschiedlich hohen Bedienraten von Deaktivierer und Markenspeicher (\diamond Tabelle 7.7) und der damit verbundenen Auslastung. Letzterer kann im vorliegenden Fall pro Millisekunde 17,678 Kunden bedienen, während der Deaktivierer in der gleichen Zeit nur 3,563 Kunden bedienen kann.

7.3 Die Anzahl der Bedieneinheiten für die Knoten des Warteschlangennetzes

Im Rahmen einer analytischen Leistungsbewertung ist mit Hilfe der Warteschlangentheorie eine Anpassung der Bedieneinheiten eines Warteschlangensystems an die zu erwartende Auftragslast möglich. Der durch die Hinzunahme von Bedieneinheiten entstehende Aufwand bei der Erstellung einer Hardware-Implementierung und der durch die Parallelisierung verursachte *Overhead* werden dabei allerdings nicht berücksichtigt. Der erreichte *Speed Up* bezieht sich nur auf die durch die im Warteschlangennetz berechenbaren mittleren Verweilzeiten $E[r]$. Er ist nicht in vollem Maße auf die Realität übertragbar und zeigt nur Tendenzen auf. Ebenfalls kann das Kosten-Nutzen-Verhältnis bei der analytischen Leistungsbewertung nicht einfließen.

7.3.1 Die Anzahl der Schaltmaschinen

In den Schaltmaschinen der PAPER-Architektur werden die Transitionsfunktionen des als PENCIL-Netz spezifizierten Protokolls und die externen Ereignisse bearbeitet. Da letztere bevorzugt zu behandeln sind, wurden die Schaltmaschinen als M/M/c-FCFS PR-Wartesystem mit zwei Prioritätsklassen modelliert.

Im Abschnitt 6.3 wurden die Ankunftsrate für die einzelnen Prioritätsklassen analysiert. Dabei stellte sich heraus, daß die Ankunftsrate λ_5^{P2} der Prioritätsklasse 2 die der Prioritätsklasse 1 dominiert, falls $t_a > \frac{1-p_{glob}}{2}$ ist. In der vorliegenden Implementierung von XTP auf dem PAPER-Emulator nimmt t_a einen Wert von 0,96019 an. Nach Tabelle 7.2 ist bekannt, daß bei diesem Wert von t_a in der aggregierten Ankunftsrate λ_5 der Schaltmaschinen nur 2% – 4% Kunden der Prioritätsklasse 1 enthalten sind. Die bevorzugte Behandlung der externen Ereignisse verzögert die Abarbeitung der Transitionen in den Schaltmaschinen somit nur unwesentlich.

Zur Ermittlung einer bestimmten Anzahl von Bedieneinheiten in einem M/M/c-Wartesystem sind *rein mathematisch* mehrere Wege denkbar.

1. Die Anzahl der Bedieneinheiten wird so gewählt, daß sich die mittlere Verweilzeit $E[r]$ bis auf eine frei wählbare Differenz ε der mittleren Bedienzeit $E[s]$ annähert.
2. Die mittlere Bedienzeit ist unabhängig von der Anzahl der Bedieneinheiten. Da die mittlere Verweilzeit die Summe aus der mittleren Wartezeit $E[q]$ und der mittleren Bedienzeit $E[s]$ ist, kann die Anzahl der Bedieneinheiten so gewählt werden, daß sich die mittlere Wartezeit bis auf eine frei wählbare Differenz ε Null nähert.
3. Ähnlich wie unter 1. und 2. kann man auch für die mittlere Kundenanzahl $E[n]$ oder die mittlere Warteschlangenlänge $E[n_q]$ vorgehen.

Wegen der deutlichen Dominanz der Prioritätsklasse 2 über die Prioritätsklasse 1 wird die Anzahl der **Schaltmaschinen** über die aggregierten Leistungsgrößen für M/M/c-FCFS PR-Wartesysteme ermittelt, welche durch die Berechnungsvorschriften für M/M/c-Wartesystem (\Leftrightarrow Abschnitt 4.3.4) berechnet werden können. Die Benutzung des globalen Speichers soll zunächst ausgeschlossen werden.

Die für die Ermittlung der Schaltmaschinenanzahl c_5 durch ein M/M/c-Wartesystem benötigten Werte sind in der folgenden Tabelle zusammengefaßt.

Ankunftsrate	λ_5	$25,119\lambda_{05}$
Bedienrate pro ms	μ_5	1,937
Auslastung	ρ_5	$\frac{12,967}{c_5}\lambda_{05}$

Damit das M/M/c-Wartesystem stationäre Zustände einnehmen kann, muß gelten: $\lambda_{05} < c_5 \cdot 0,077$. Daher wird als Ausgangsgröße zunächst $\lambda_{05} \leq 0,07$ gewählt. Dabei zeigt sich, daß die aggregierte mittlere Wartezeit $E_5[r]$ der Schaltmaschinen beim Einsatz von drei Ausführungseinheiten einen Wert nahe 0 einnimmt (\Leftrightarrow Abbildung 7.5).

Es ergeben sich für die mittlere Verweilzeit $E_5[r]$ die in der Tabelle 7.10 aufgeführten *Speed Ups*.

λ_{05} pro ms	Bedieneinheiten	Speed Up	Auslastung ρ_5
0,0175	3	1,29	7%
0,0350	4	1,83	11%
0,0525	5	3,13	13%
0,0700	6	10,83	18%

Tabelle 7.10: Der erreichte *Speed Up* beim Einsatz mehrerer Bedieneinheiten

Die Abbildung 7.6 verdeutlicht diesen Zusammenhang.

Bei der bisherigen Betrachtung der Schaltmaschinen war die Benutzung des globalen Speichers ausgeschlossen. Im Abschnitt 6.2.4 wurde schon auf dessen besondere Form der Einbindung in das Warteschlangennetz der PAPER-Architektur eingegangen. Die Bedienrate der Schaltmaschinen muß in Abhängigkeit der den globalen Speicher nutzenden Kunden korrigiert werden. Benutzen die mit der Rate $p_{45}\lambda_4 + \lambda_{05}$ an den Schaltmaschinen ankommenden Kunden mit der Wahrscheinlichkeit p_{glob} den globalen Speicher, so ist die Bedienrate nach Gleichung (6.3) um

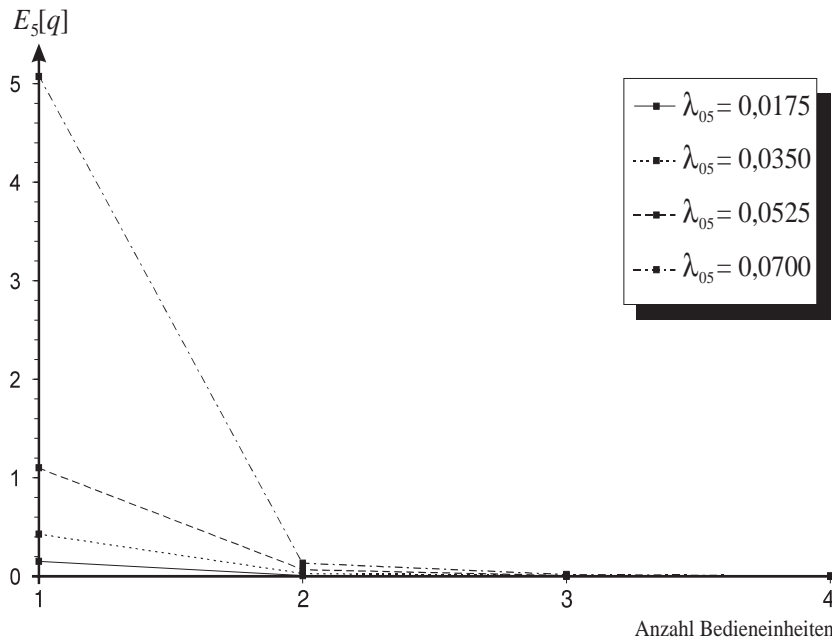


Abbildung 7.5: Die mittlere Wartezeit beim Einsatz mehrerer Bedieneinheiten

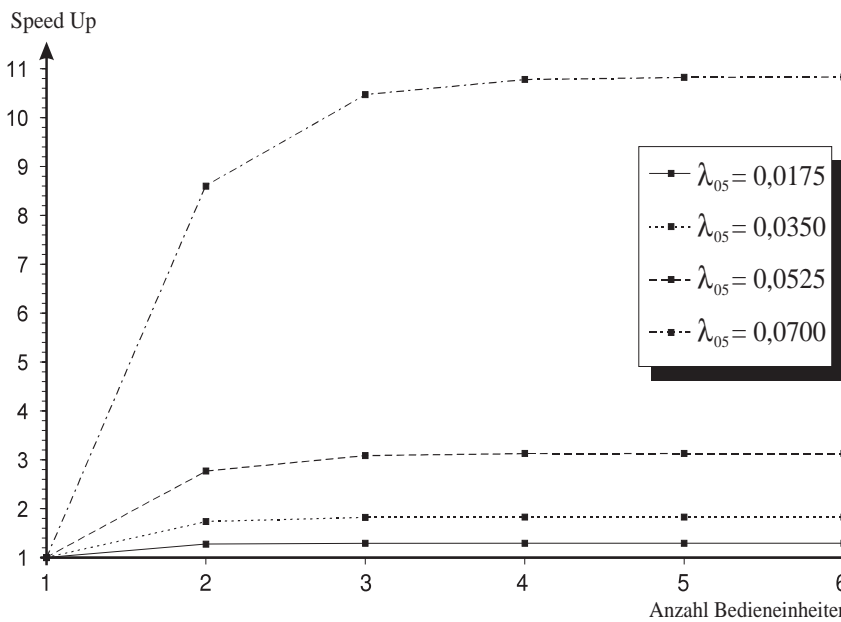


Abbildung 7.6: Der *Speed Up* beim Einsatz mehrerer Bedieneinheiten

den Faktor $(1 + p_{glob})$ zu korrigieren. Betrachtet man anschließend das gesamte Warteschlangennetz, so ist die mittlere Verweilzeit eines Kunden in den Schaltmaschinen auch um diesen Faktor zu korrigieren, da er die Schaltmaschinen im Mittel $(1 + p_{glob})$ -mal durchläuft. In der PAPER-Architektur erfolgt das Schalten einer Transition oder eines externen Ereignisses in den Schaltmaschinen nämlich nur einmal. Durch die Benutzung des globalen Speichers erhöhen sich die Bedien- und Ankunftsrate der Schaltmaschinen in gleichem Maße. Im Hinblick auf das Gesamtverhalten des Warteschlangennetzes hat die Benutzung des globalen Speichers daher keine Auswirkung auf die Anzahl der einzusetzenden Ausführungseinheiten und damit auf das Zeitverhalten an den Schaltmaschinen.

Die Anzahl der in der PAPER-Architektur einzusetzenden Schaltmaschinen ist in erster Linie von dem zu erwartenden Lastaufkommen abhängig. Mit dem vorgestellten Verfahren zur Minimierung der mittleren Wartezeit läßt sich mit Hilfe der Warteschlangentheorie die Anzahl der einzusetzenden Schaltmaschinen auf die erwartete Kundenlast abstimmen. Eine Analyse der Tabelle 7.7 zeigt, daß die Schaltmaschinen für $c_5 = 1$ die externe Ankunftsrate λ_{05} für das Warteschlangennetz begrenzen. Schon der Einsatz einer weiteren Bedieneinheit führt zu einer Erhöhung der möglichen Werte von λ_{05} . Durch den Einsatz mehrerer Bedieneinheiten wird eine schnellere Kundenverarbeitung erreicht, was für die PAPER-Architektur einen höheren Grad an Parallelisierung und eine schnellere Verarbeitung von Protokollaktionen bedeutet.

7.3.2 Die Bedieneinheiten weiterer Knoten

Beim Einsatz mehrerer Bedieneinheiten für das M/M/c-Wartesystem der Schaltmaschinen wird λ_{05} durch die Knoten der Kontrolleinheit beschränkt. Diese sind alle als M/M/1-Wartesystem modelliert. In diesem Abschnitt wird die Steigerung der Leistungsfähigkeit der PAPER-Architektur durch den *mehrfachen* Einsatz von Komponenten der Kontrolleinheit untersucht. Das umfaßt zum einen die Möglichkeit eine höhere Last zu verarbeiten und durch Parallelisierung in den einzelnen Komponenten eine schnellere Verarbeitungsgeschwindigkeit zu erreichen. Wird beispielsweise zum Bereitstellen einer aktivierbaren Transition mehr Zeit benötigt als das Schalten dieser Transition benötigt, so können zwar mehrere Transitionen in den Schaltmaschinen parallel ausgeführt werden, die Protokollverarbeitung wird aber insgesamt verzögert. Zur Analyse dieses Sachverhalts wird das Zeitverhalten der einzelnen Knoten des Warteschlangennetzes betrachtet. Dabei wird von drei Schaltmaschinen ausgegangen. Die maximale Rate für λ_{05} ist 0,1 pro ms (\Leftrightarrow Tabelle 7.7).

Knoten	λ_i	Type	Auslastung	$E_i[r]$	$E_i[q]$	$E_i[s]$
Deaktivierer	2,512	M/M/1	70,5%	0,951	0,671	0,280
Auswerter	54,007	M/M/1	96,6%	0,529	0,511	0,018
Sperreinheit	5,544	M/M/1	57,9%	0,248	0,144	0,104
Markenspeicher	4,924	M/M/1	27,9%	0,078	0,022	0,056
Schaltmaschinen	2,512	M/M/3	43,2%	0,568	0,051	0,517

Tabelle 7.11: Das Zeitverhalten des Warteschlangennetzes für $\lambda_{05} = 0,1$ (Zeiten in ms)

Die Tabelle 7.11 zeigt für die Knoten der Kontrolleinheit hohe mittlere Wartezeiten $E_i[q]$ im Vergleich zu ihren Bedienzeiten $E_i[s]$. Dies wird durch die hohe Auslastung der Knoten hervorgerufen und hat seine Ursache in zwei Gründen. Zum einen hat der Deaktivierer eine lange Bedienzeit. Zum anderen hat der Auswerter eine sehr hohe Ankunftsrate, die durch den hohen Konnektivitätsgrad des PENCIL-Netzes von XTP hervorgerufen wird. Die Bedienung von Kunden durch den Deaktivierer und den Auswerter wird durch hohe Wartezeiten verzögert. Im weiteren Verlauf dieses Abschnitts werden diese Komponenten innerhalb der PAPER-Architektur mehrfach eingesetzt. Dies führt zu einer weiteren Parallelisierung bei der Bereitstellung aktivierbarer Tran-

sitionen. Durch die Form der Modellierung des Warteschlangennetzes ist gewährleistet, daß diese Parallelisierung die notwendige Synchronisationsfunktion der Sperreinheit nicht beeinflusst.

Der **Deaktivierer** nimmt die Transitionen, die in der Ausführungseinheit geschaltet haben, und die externen Ereignisse in Form von Deaktivierungspaketen entgegen. Er leitet die dadurch potentiell schaltfähig gewordenen Transitionen an den Auswerter weiter und entsperrt die vom Aktivierer zum Schalten der zu bearbeitenden Transition gesperrten Transitionen. Das Entsperren und Sperren von Transitionen wird durch den Sperrvektor synchronisiert. Dadurch können in der PAPER-Architektur mehrere Deaktivierer eingesetzt werden.

Im Warteschlangennetz wird der Deaktivierer nun als M/M/c-Wartesystem betrachtet. Zur Abstimmung der Anzahl der einzusetzenden Bedieneinheiten auf die Kundenlast wird die mittlere Wartezeit $E_1[q]$ für verschiedene λ_{05} betrachtet. Damit das M/M/c-Wartesystem des Deaktivierers stationäre Zustände einnehmen kann, muß $\lambda_{05} < 0,142$ gelten. Die Abbildung 7.7 zeigt den erreichten *Speed Up* beim Einsatz mehrerer Bedieneinheiten für verschiedene Kundenlasten.

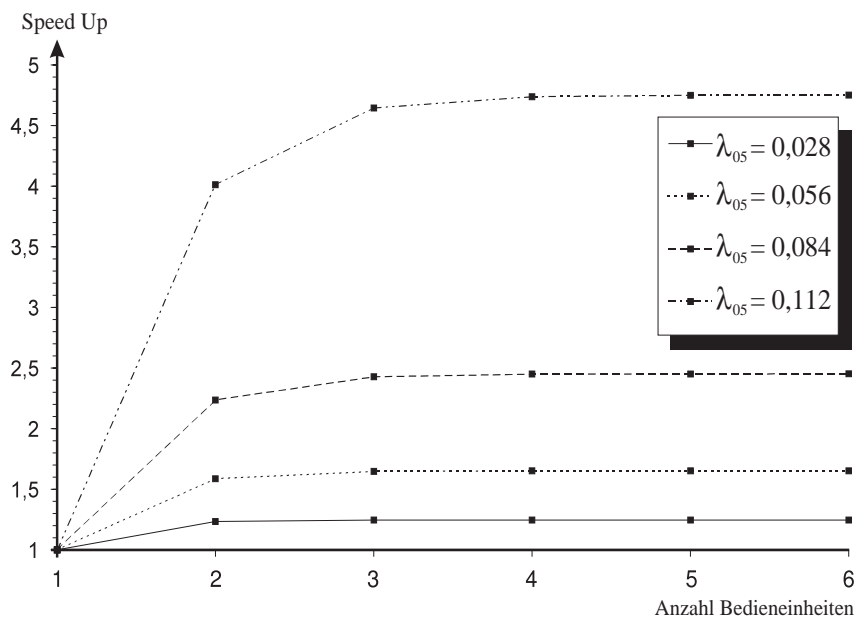


Abbildung 7.7: Der *Speed Up* für die mittlere Wartezeit des Deaktivierers

Der erreichte *Speed Up* der mittleren Wartezeit $E_1[q]$ für die Kunden des Deaktivierers ist in der Tabelle 7.12 zusammengefaßt.

λ_{05} pro ms	Bedieneinheiten	Speed Up	Auslastung ρ_1
0,028	3	1,25	6%
0,056	4	1,65	9%
0,084	5	2,45	11%
0,112	6	4,75	13%

Tabelle 7.12: Der erreichte *Speed Up* beim Einsatz mehrerer Bedieneinheiten

Der Einsatz von fünf Deaktivierern bei der im Warteschlangennetz maximal möglichen Rate von $\lambda_{05} < 0,104$ (\diamond Tabelle 7.7) führt somit zu einer ca. 3,4-fach schnelleren Bearbeitung der Kunden.

Der **Auswerter** wertet aus der aktuellen Markierung des PENCIL-Netzes die Schaltbedingungen der potentiell schaltfähigen Transitionen aus und ermittelt somit die Protokollaktionen, die als nächstes auszuführen sind. Dabei erfolgt kein Zugriff auf eine Komponente der PAPER-Architektur, die eine Zugriffssynchronisation erfordert. Somit läßt sich auch der Auswerter in der PAPER-Architektur mehrfach einsetzen und im Warteschlangennetz als M/M/c-Wartesystem modellieren.

Wie beim Deaktivierer wird die Anzahl der einzusetzenden Bedieneinheiten durch eine Betrachtung der mittleren Wartezeit $E_2[q]$ für verschiedene λ_{05} ermittelt. Dabei ergeben sich der in der Abbildung 7.8 und Tabelle 7.13 dargestellten *Speed Up*.

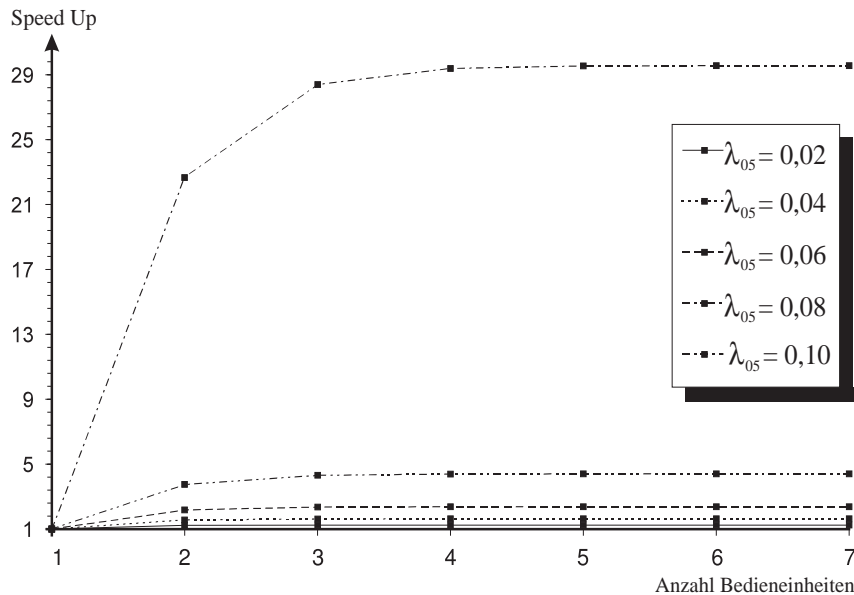


Abbildung 7.8: Der *Speed Up* für die mittlere Wartezeit des Auswerter

λ_{05} pro ms	Bedieneinheiten	Speed Up	Auslastung ρ_2
0,02	3	1,24	6%
0,04	4	1,63	9%
0,06	5	2,38	11%
0,08	6	4,40	12%
0,10	7	29,57	13%

Tabelle 7.13: Der erreichte *Speed Up* beim Einsatz mehrerer Bedieneinheiten

Bisher wurde die Höhe der externen Ankunftsrate des Warteschlangennetzes durch den Auswerter bestimmt. Durch den Einsatz mehrerer Auswerter ist nun ein höherer Wert für λ_{05} möglich. Bei dem bisher gültigen Wert von $\lambda_{05} = 0,1$ führt der Einsatz von sieben Auswertern zu einem *Speed Up* der mittleren Verweilzeit $E_2[r]$ von 29,57.

Die Analyse der Knoten des Warteschlangennetzes für die Schaltmaschinen, den Deaktivierer und den Auswerter haben ergeben, daß durch eine Anpassung der Anzahl von Bedieneinheiten an die zu erwartende Kundenlast eine Verkürzung der mittleren Wartezeit möglich ist. Auf diese Weise wird eine weitere Parallelisierung bei der Bereitstellung aktivierbarer Transitionen erreicht, was zu einer Beschleunigung bei der Protokollverarbeitung führen kann.

Es sei noch einmal darauf hingewiesen, daß der auf der Basis der Warteschlangentheorie erreichte *Speed Up* für die jeweiligen Knoten *keinen* durch Parallelisierung entstehenden Overhead berücksichtigen und somit alle Aussagen tendenziell sind.

7.4 Das Verhältnis von Aktivierungs- und Schaltzeit

Bisher sind die Komponenten der PAPER-Architektur nur getrennt voneinander untersucht worden. Die Leistungsfähigkeit der PAPER-Architektur hängt nicht alleine nur von der der einzelnen Komponenten ab, sondern vielmehr von dem Zusammenwirken von Kontroll- und Ausführungseinheit. Dieses Zusammenwirken wird bestimmt durch die *Aktivierungs-* und die *Schaltzeit*. Die Aktivierungszeit ist die Zeit, die von der Kontrolleinheit zur Aktivierung *einer* Transition durchschnittlich benötigt wird. Die Schaltzeit ist die Zeit, die eine aktivierte Transition in der Ausführungseinheit im Mittel verbringt. Die Betrachtung dieser Zeiten ist Gegenstand dieses Abschnitts.

Im Abschnitt 7.3 wurde aufgezeigt, daß sich in der PAPER-Architektur die Transitionsverarbeitung neben den Schaltmaschinen auch durch den Einsatz mehrerer Deaktivierer und Auswerter beschleunigen läßt. Im Warteschlangennetz der PAPER-Architektur werden Deaktivierer und Auswerter dann als M/M/c-Wartesystem modelliert. Die Benutzung mehrerer Bedieneinheiten führt zu einer Verringerung der Auslastung ρ_i in den betroffenen Knoten. Gemäß der Tabelle 7.7 muß zum Erhalt der Local-Balance-Eigenschaft des Warteschlangennetzes $\lambda_{05} < 0,173$ gelten.

7.4.1 Die Schaltzeit

Das Schalten einer Transition des als PENCIL-Netz formalisierten Protokolls wird im Warteschlangennetz durch die Bedienung in den Knoten Markenspeicher und Schaltmaschinen modelliert. Dabei durchlaufen die Kunden den Markenspeicher zweimal. Einmal während der Aktivierungsphase einer Transition und einmal während der Deaktivierungsphase. Die *Schaltzeit* läßt sich durch Addition der mittleren Verweilzeiten in den einzelnen Knoten berechnen. Die hier betrachteten Kunden sind an den Schaltmaschinen Kunden der Prioritätsklasse 2. Wird der globale Speicher zunächst ausgeschlossen, so läßt sich die Schaltzeit durch

$$\tau_s = 2 \cdot E_4[r] + E_5^{P2}[r] \quad (7.5)$$

berechnen.

Durch die Anzahl der Schaltmaschinen ist es möglich die mittlere Verweilzeit $E_5^{P2}[r]$ eines Kunden auf die mittlere Bedienzeit $E_5[s]$ zu reduzieren. Bei der im Warteschlangennetz maximal möglichen Rate von $\lambda_{05} = 0,17$ müssen mindestens drei Bedieneinheiten für die Schaltmaschinen verwendet werden. In der Tabelle 7.14 sind die bei der Untersuchung ermittelten Ergebnisse für τ_s aufgeführt.

Die Schaltzeit läßt sich bei der maximalen Ankunftsrate für die externen Ereignisse durch $\tau_s = 1,1 \text{ ms}$ nach oben begrenzen. Bei höheren Raten von λ_{05} steigt die Auslastung des Marken-

Bedieneinheiten		Werte für λ_{05}					
c_4	c_5	0,050	0,075	0,100	0,125	0,150	0,170
1	3	0,655	0,682	0,726	0,780	0,924	1,099
1	4	0,648	0,663	0,682	0,711	0,753	0,800
1	5	0,648	0,660	0,675	0,694	0,720	0,747
1	6	0,648	0,659	0,673	0,691	0,713	0,735
1	7	0,648	0,659	0,673	0,690	0,711	0,732

Tabelle 7.14: Die *Schaltzeit* τ_s einer Transition (Zeiten in *ms*)

speichers und der Schaltmaschinen an. Das hat eine Erhöhung der mittleren Verweilzeit in den Knoten zur Folge. Durch eine Anpassung der Bedieneinheiten für die Schaltmaschinen an das erhöhte Lastaufkommen läßt sich diese wieder reduzieren. Das ist für den Markenspeicher nicht möglich. Die Zugriffe auf diesen müssen synchronisiert werden.

Benutzen Transitionen den globalen Speicher, so führt dies zwangsläufig zu einer Erhöhung der Schaltzeit. Im Abschnitt 6.2.4 wurde ausgeführt, daß die Bedienrate der Schaltmaschinen um den Faktor $1 + p_{glob}$ zu korrigieren ist, wenn die Kunden mit einer Wahrscheinlichkeit von p_{glob} den globalen Speicher benutzen. Das führt zu einer Verkürzung der aggregierten mittleren Verweilzeit und den mittleren Verweilzeiten für die beiden Prioritätsklassen. Da die Kunden die Schaltmaschinen $(1 + p_{glob})$ -mal durchlaufen, verbringen sie nach der Konstruktion des Warteschlangennetzes *insgesamt* genauso viel Zeit in den Schaltmaschinen, wie ohne Benutzung des globalen Speichers. Die Berechnung der Schaltzeit unter Verwendung des globalen Speichers erfolgt durch

$$\tau_s = 2 \cdot E_4[r] + E_5^{P^2}[r] + E_6[r] \quad (7.6)$$

Für die Benutzungshäufigkeit und Zugriffsdauer auf den globalen Speicher liegen hinsichtlich der Implementierung von XTP auf dem PAPER-Emulator keine Angaben vor. Der globale Speicher und die Zugriffe über den globalen Bus auf diesen werden im Warteschlangennetz als M/M/1-Wartesystem modelliert. Die mittlere Verweilzeit eines M/M/1-Wartesystems hat die Eigenschaft, daß sie bei hohen Auslastungen stark ansteigt. Eine hohe Auslastung wird entweder durch viele Kunden mit kurzen Bedienzeiten oder durch wenige Kunden mit langen Bedienzeiten hervorgerufen. Ähnlich ist auch das Zugriffsverhalten der Schaltmaschinen auf den globalen Speicher anzusehen. Nimmt der Speicherzugriff einen längeren Zeitraum in Anspruch, so ist der Bus in diesem Zeitraum für andere Schaltmaschinen, die auch auf den globalen Speicher zugreifen wollen, blockiert und das Abarbeiten der Transitionfunktion verzögert sich. Erfolgen dagegen viele kurze Zugriffe, wird die Verzögerung durch die Häufigkeit der Zugriffe hervorgerufen. Im Hinblick auf eine möglichst verzögerungsfreie parallele Bearbeitung der Transitionfunktion durch die Schaltmaschinen ist das PENCIL-Netz eines Kommunikationsprotokolls so in PENCIL/C zu implementieren, daß für die Benutzung des globalen Speichers so wenig Bedarf wie möglich entsteht.

7.4.2 Die Aktivierungszeit

Das Bereitstellen von aktivierbaren Transitionen durch die Kontrolleinheit erfolgt im Warteschlangennetz der PAPER-Architektur durch die Knoten Deaktivierer, Auswerter und Sperreinheit. Nach den ermittelten Werten bei der Emulation von PAPER mittels XTP hat die Bearbeitung einer Transition oder eines externen Ereignisses durch den Deaktivierer zur Folge, daß inzidenzbedingt im Mittel 21,5 Transitionen auf ihre Schaltfähigkeit zu prüfen sind. Von diesen werden $t_s \stackrel{(5.17)}{=} \gamma p_{23} = 1,247$ Transitionen als schaltfähig erkannt. Von den schaltfähigen Transitionen werden 77% aktiviert und der Ausführungseinheit zugeführt. Die Zeit, die dieser Vorgang zur Aktivierung *einer* Transition benötigt, wird im folgenden *Aktivierungszeit* genannt.

Zur Berechnung der Aktivierungszeit wird von dem im Kapitel 5 konstruierten Warteschlangennetz ausgegangen. Neben der Anzahl der Transitionen müssen auch die flußbedingten Abhängigkeiten von Kunden der Menge \mathcal{K} im Warteschlangennetz berücksichtigt werden. So greifen Kunden des Deaktivierers mit einer Wahrscheinlichkeit von p_{13} auf die Sperreinheit zu. Dadurch wird das Entsperren von Transitionen modelliert. Die Aktivierungszeit τ_a für eine Transition läßt sich dann durch die mittleren Verweilzeiten für die Knoten der Kontrolleinheit berechnen.

$$\begin{aligned} \tau_a &= \frac{E_1[r] + p_{13}E_3[r] + \gamma E_2[r] + t_s E_3[r]}{t_a} \\ &= \frac{E_1[r] + \gamma E_2[r] + (p_{13} + t_s)E_3[r]}{t_a} \end{aligned} \quad (7.7)$$

Für eine Untersuchung von τ_a wird zunächst davon ausgegangen, daß alle Knoten der Kontrolleinheit vom Typ M/M/1 sind. Anschließend wird das Verhalten der Schaltzeit für den Einsatz von mehreren Deaktivierern und Auswertern untersucht. Dabei wird auf die im Abschnitt 7.3 ermittelten Werte für die Anzahl der Bedieneinheiten zurückgegriffen. Die externe Ankunftsrate wird dabei im zulässigen Bereich von $\lambda_{05} \leq 0,17$ variiert. Es ist zu beachten, daß sich die vom Auswerter insgesamt benötigte Zeit bei Benutzung mehrerer Bedieneinheiten durch die entstehende Parallelisierung von $\gamma E_2[r]$ auf $\frac{\gamma}{c_2} E_2[r]$ reduziert. In der Tabelle 7.15 sind die bei der Untersuchung ermittelten Ergebnisse für τ_a aufgeführt.

Bedieneinheiten			Werte für λ_{05}					
c_1	c_2	c_3	0,050	0,075	0,100	0,125	0,150	0,170
1	1	1	1,564	2,500	13,405	—	—	—
6	1	1	1,405	2,172	12,706	—	—	—
1	7	1	0,847	1,102	1,619	3,387	—	—
6	7	1	0,688	0,774	0,920	1,220	2,181	16,012

Tabelle 7.15: Die *Aktivierungszeit* τ_a einer Transition (Zeiten in ms)

Die Betrachtung der Aktivierungszeit zeigt, daß eine Beschleunigung im vorliegenden Fall hauptsächlich durch den Einsatz mehrerer Auswerter erreicht wird. Die Ursache liegt in der im Vergleich zu den anderen Knoten sehr hohen Ankunftsrate. Diese wird durch die Vielzahl der potentiell schaltfähigen Transitionen hervorgerufen. Damit kommt dem Auswerter im Hin-

blick auf die Aktivierungszeit eine zentrale Bedeutung zu. Das muß jedoch nicht immer der Fall sein. In PENCIL-Netzen, in denen die Anzahl der indirekt inzidenten Transitionen geringer ist, wird der Auswerter nicht so sehr ausgelastet wie im vorliegenden Fall. Den Engpaß für die Aktivierungszeit im Warteschlangennetz der PAPER-Architektur ist die Sperreinheit, was sich in den langen Aktivierungszeiten bei höheren Auslastungen zeigt (\diamond Tabelle 7.15). Dieser Engpaß wird durch die notwendige Synchronisation beim Zugriff auf den Sperrvektor zum Sperren und Entsperren von Transitionen hervorgerufen und kann eine Beschleunigung durch den Einsatz mehrerer Auswerter und Deaktivierer zunichte machen.

7.4.3 Vergleich und Bewertung von Aktivierungs- und Schaltzeit

Ein Vergleich der beiden Zeiten zeigt, daß die Aktivierungszeit bei geringen Raten von λ_{05} nur etwas größer ist als die Schaltzeit.

Bedieneinheiten					Werte für λ_{05}					
c_1	c_2	c_3	c_4	c_5	0,050	0,075	0,100	0,125	0,150	0,170
1	1	1	1	3	2,39	3,67	18,45	—	—	—
1	1	1	1	7	2,42	3,79	19,92	—	—	—
6	7	1	1	3	1,05	1,14	1,27	1,53	2,36	14,57
6	7	1	1	7	1,06	1,17	1,37	1,77	3,06	21,87

Tabelle 7.16: Das Verhältnis von Aktivierungszeit zu Schaltzeit durch $\frac{\tau_a}{\tau_s}$

In einem Warteschlangennetz, bei dem alle Knoten der Kontrolleinheit vom Typ M/M/1 sind, wirkt sich eine Erhöhung der Bedieneinheiten für die Schaltmaschinen kaum aus. Diese haben dann wegen der langen Aktivierungszeit eine nur geringe Auslastung. Erst die Erhöhung der Bedieneinheiten für die Kontrolleinheit führt zu einer deutlichen Verbesserung des Verhältnisses von Aktivierungs- zu Schaltzeit. Die langen Aktivierungszeiten werden durch die hohe Konnektivität des PENCIL-Netzes für XTP hervorgerufen. Geht man von einem weniger konnektiven PENCIL-Netz aus, so sind wegen der geringeren Auslastung des Auswerter kürzere Aktivierungszeiten zu erwarten. Der günstigste Fall für die Protokollausführung ist der Fall, bei dem die Aktivierungszeit kürzer als die Schaltzeit ist, d.h. $\frac{\tau_a}{\tau_s} < 1$. Dabei wird zum einen von der Möglichkeit der parallelen Ausführung von Protokollaktionen durch die Transitionsfunktionen Gebrauch gemacht, und der durch das Bereitstellen von aktivierbaren Transitionen entstehende zusätzliche Aufwand wird klein gehalten. Das führt zu der Schlußfolgerung, daß die Leistungsfähigkeit der PAPER-Architektur nicht nur von einer Abstimmung der Anzahl einzusetzender Komponenten auf die erwartete Last abhängt, sondern auch von der Granularität des PENCIL-Netzes für das verwendete Kommunikationsprotokoll. Die Granularität hat dabei Einfluß auf die Möglichkeit der parallelen Verarbeitung von Transitionen. Werden die Protokollaktionen in eine zu große Anzahl von Transitionen zerlegt, so kommt es zu einer Erhöhung des Aktivierungsaufwandes, der sich in zu langen Aktivierungszeiten niederschlagen kann.

7.5 Die Umlaufzeit einer Transition

Das Aktivieren, Schalten und Deaktivieren einer Transition des als PENCIL-Netz spezifizierten Protokolls wird durch das sequentielle Durchlaufen eines Kunden der Menge \mathcal{T} durch die Knoten des Warteschlangennetzes modelliert. Die Summe der mittleren Verweilzeiten $E_i[r]$ ist die *Umlaufzeit*.

Die auf ihre Schaltfähigkeit zu prüfenden Transitionen werden dem Warteschlangennetz als Kunden der Menge \mathcal{T} über die Quelle Q_1 zugeführt. Sie durchlaufen den Auswerter, die Sperreinheit und den Markenspeicher. Das Ausführen der Transitionfunktion findet dann durch die Bedienung des Kunden in den Schaltmaschinen statt. Die Änderungen der Markierung des PENCIL-Netzes werden durch ein erneutes Durchlaufen des Markenspeichers modelliert. Die Deaktivierung der Transition und das Freigeben von gesperrten Transitionen erfolgt durch die Bedienung im Deaktivierer und der Sperreinheit. Anschließend verläßt der Kunde das Warteschlangennetz durch die Senke S_1 . Die Umlaufzeit τ_u läßt sich somit durch

$$\begin{aligned} \tau_u &= E_2[r] + E_3[r] + 2 \cdot E_4[r] + E_5^{P2}[r] + E_1[r] + E_3[r] \\ &\stackrel{(7.5)}{=} E_1[r] + E_2[r] + 2 \cdot E_3[r] + \tau_s \end{aligned}$$

berechnen.

Werden die aus der XTP-Implementierung auf dem PAPER-Emulator bekannten Werte für das Warteschlangennetz verwendet, so ergeben sich unter Einbeziehung der im Abschnitt 7.3 ermittelten Bedieneinheitenanzahl die in Tabelle 7.17 aufgelisteten mittleren Verweilzeiten eines Kunden in der Kontrolleinheit. Die externe Ankunftsrate wird für den im Warteschlangennetz gültigen Bereich $\lambda_{05} \leq 0,17$ betrachtet.

Bedieneinheiten			Werte für λ_{05}					
c_1	c_2	c_3	0,050	0,075	0,100	0,125	0,150	0,170
1	1	1	0,762	1,030	1,977	—	—	—
6	1	1	0,609	0,715	1,306	—	—	—
1	7	1	0,745	0,983	1,466	—	—	—
6	7	1	0,593	0,668	0,795	1,056	1,891	13,925

Tabelle 7.17: Die *mittlere Verweilzeit* eines Kunden in der Kontrolleinheit (Zeiten in *ms*)

Die Zeit, die ein Kunde in der Ausführungseinheit verbringt (τ_s), wurde schon im vorangegangenen Abschnitt ermittelt. Die Addition dieser Werte (\diamond Tabelle 7.14) und die oben berechneten Zeiten für die Kontrolleinheit ergibt die Umlaufzeit τ_u eines Kunden im Warteschlangennetz (\diamond Tabelle 7.18).

Eine Analyse der Zusammensetzung von τ_u zeigt, daß in den meisten Fällen die Schaltzeit τ_s kleiner ist als die Zeit, die der Kunde in der Kontrolleinheit verbringt (\diamond Tabelle 7.19). Die Umlaufzeit ändert sich bei der Hinzunahme weiterer Bedieneinheiten für die Schaltmaschinen kaum. Erst eine Erhöhung der Bedieneinheiten für die Komponenten der Kontrolleinheit führt

Bedieneinheiten					Werte für λ_{05}					
c_1	c_2	c_3	c_4	c_5	0,050	0,075	0,100	0,125	0,150	0,170
1	1	1	1	3	1,417	1,712	2,703	—	—	—
1	1	1	1	7	1,410	1,689	2,650	—	—	—
6	7	1	1	3	1,247	1,350	1,522	1,855	2,815	15,024
6	7	1	1	7	1,240	1,327	1,468	1,746	2,602	14,657

Tabelle 7.18: Die Umlaufzeit τ_u in ms

zu einer Verkürzung der Umlaufzeit. Die Umlaufzeit wird im vorliegenden Fall somit durch die Zeit dominiert, die der Kunde in der Kontrolleinheit verbringt. Das spiegelt sich auch in den Prozentzahlen der Schaltzeit in Bezug auf die Umlaufzeit wieder.

Bedieneinheiten					Werte für λ_{05}					
c_1	c_2	c_3	c_4	c_5	0,050	0,075	0,100	0,125	0,150	0,170
1	1	1	1	3	46,20	39,83	26,87	—	—	—
1	1	1	1	7	45,94	39,03	25,40	—	—	—
6	7	1	1	3	52,48	50,51	47,75	43,10	32,82	7,31
6	7	1	1	7	52,21	49,67	45,84	39,52	27,32	5,00

Tabelle 7.19: Der prozentuale Anteil der Schaltzeit τ_s an der Umlaufzeit τ_u

Die Ursache für die Dominanz der Kontrolleinheit sind die Wartezeiten an den einzelnen Knoten. Dies gilt besonders für den Auswerter und bei einem hohen Kundenaufkommen für die Sperreinheit. Die lange mittlere Wartezeit für Kunden am Auswerter hat seine Ursache in der Konnektivität des PENCIL-Netzes von XTP. Bei einem weniger hohen Grad an Konnektivität führt das zu einer Verkürzung der Zeit, die ein Kunde in der Kontrolleinheit verbringt. Ähnlich wie im Abschnitt 7.4.2 wirkt sich bei einem hohen Kundenaufkommen an der Sperreinheit diese verzögernd auf die Umlaufzeit aus. Durch ihre Synchronisationsfunktion bei der Modellierung der Sperrvektorzugriffe in der PAPER-Architektur kann sie eine Verkürzung der Umlaufzeit, welche durch Hinzunahme von Bedieneinheiten für die Knoten Deaktivierer und Auswerter erreicht werden kann, ins Gegenteil umkehren.

7.6 Zusammenfassung der Analyse

In diesem Kapitel wurden die bei einer Teilimplementierung von XTP auf dem PAPER-Emulator ermittelten Zeiten für die Transitionsbearbeitung auf das Warteschlangenmodell der PAPER-Architektur angewendet. Dabei hat sich durch die Analyse der Ankunftsdaten im Warteschlangenmodell gezeigt, daß die Auftragslast der einzelnen Komponenten der PAPER-Architektur von

1. der Konnektivität des PENCIL-Netzes,
2. der Häufigkeit des Auftretens externer Ereignisse

abhängt. Durch eine Analyse der erwarteten Auftragslast und dem modularen Aufbau der PAPER-Architektur läßt sich die Anzahl der jeweiligen Komponenten auf die erwartete Auftragslast abstimmen. So kann durch den Einsatz mehrerer Schaltmaschinen, Deaktivierer und Auswerter eine Geschwindigkeitssteigerung bei der Transitionsbearbeitung erreicht werden.

Die Analyse des gesamten Warteschlangennetzes hat die Abhängigkeit der Leistungsfähigkeit vom Zusammenwirken der Kontroll- und Schalteinheit aufgezeigt. Dazu wurden die für die Aktivierung und das Schalten der Transitionen benötigten Zeiten untersucht. Diese Zeiten werden sowohl von der Konnektivität als auch von der Granularität des PENCIL-Netzes beeinflusst. Der bei stark konnektiven Netzen benötigte Zeitaufwand für die Aktivierung von Transitionen kann die Möglichkeit der parallelen Ausführung durch die Schaltmaschinen einschränken oder sogar unmöglich machen. In ähnlicher Weise wirkt sich die Granularität des PENCIL-Netzes auf die Leistungsfähigkeit der PAPER-Architektur aus. Werden die Protokollaktionen in zuviele Transitionen mit kurzen Transitionsfunktionen zerlegt, so erhöht sich der Aufwand für die Aktivierung, während das Schalten einer Transition nur wenig Zeit in Anspruch nimmt.

Zusammenfassend gibt die folgende Auflistung einen Überblick über die wichtigsten Kriterien zur Förderung der Leistungsfähigkeit der PAPER-Architektur:

- eine möglichst geringe Konnektivität des PENCIL-Netzes
- die Berücksichtigung der Schaltzeit bei der Zerlegung der Protokollaktionen in Transitionen
- eine Abstimmung der Anzahl von einzusetzenden Schaltmaschinen auf die Anzahl der parallel ausführbaren Transitionen und das allgemein erwartete Lastaufkommen
- der mehrfache Einsatz von Komponenten der Kontrolleinheit zur Beschleunigung des Aktivierungsvorgangs

8. Zusammenfassung und Ausblick

Das Ziel dieser Diplomarbeit war die analytische Leistungsbewertung einer parallelen Controller-Architektur für Hochgeschwindigkeitsprotokolle.

Zunächst wurden die Anwendungsgebiete und Methoden der Leistungsbewertung vorgestellt. Dabei wurde der Schwerpunkt auf die Modellbildungstechniken gelegt, zu denen die analytische Leistungsbewertung gehört. In diesem Zusammenhang wurde festgestellt, daß der Modellerstellung bei der analytischen Leistungsbewertung eine zentrale Bedeutung zukommt.

Im dritten Kapitel wurde die parallele Controller-Architektur PAPER beschrieben. Die Beschreibung umfaßt die Einbettung in einem Gesamtkonzept zur Entwicklung eines parallelen Kommunikations-Controllers, die Erläuterung des funktionalen Konzepts und die Beschreibung der einzelnen Komponenten.

Eine Möglichkeit zur analytischen Leistungsbewertung ist die Modellwelt der Warteschlangentheorie, die in der vorliegenden Arbeit verwendet wurde. Zum besseren Verständnis der Wartesysteme wurden zunächst einige stochastische Prozesse und deren Eigenschaften erläutert. Dann folgte eine Beschreibung von elementaren Wartesystemen mit exponentiell verteilten Ankunfts- und Bedienraten. Da sich die PAPER-Architektur aus mehreren Komponenten zusammensetzt, ist eine Modellierung als Warteschlangennetz notwendig. Diese wurden am Ende des vierten Kapitels eingeführt.

Das fünfte Kapitel beschäftigte sich mit der Erstellung eines Warteschlangenmodells für die PAPER-Architektur. Dazu wurde zuerst eine Kundenmenge definiert, die sich aus den Transitionen eines PENCIL-Netzes und den externen Ereignissen zusammensetzt. Gemäß der funktionalen Konzeption der PAPER-Architektur wurden dann die Warteschlangennetze für die Kontroll- und die Ausführungseinheit konstruiert. Dazu mußten zunächst die für die Leistungsbewertung wichtigen Komponenten und ihre Interaktionen erfaßt werden. Durch letztere wird das Ablaufgeschehen innerhalb der PAPER-Architektur beschrieben, welches durch die Raten für die Ankunftsprozesse und die Übergangswahrscheinlichkeiten an den Knoten des Warteschlangennetzes modelliert wurden.

Mit Hilfe des Warteschlangennetzes und den hergeleiteten Berechnungsvorschriften für die Ankunftsrate und die Übergangswahrscheinlichkeiten wurde dann das Kundenverhalten im Warteschlangennetz analysiert. Dabei wurde festgestellt, daß die Kundenlast, d.h. die von der PAPER-Architektur zu verarbeitende Transitionsmenge, an den einzelnen Knoten neben der Ankunftsrate von externen Ereignissen auch von der Struktur des als PENCIL-Netz formalisierten Protokolls abhängig ist.

Eine quantitative Bewertung wurde im Kapitel 7 durchgeführt. Dabei wurden Werte verwendet, die bei einer Teilimplementierung des Kommunikationsprotokolls XTP auf einem PAPER-Emulator ermittelt wurden. Diese Werte wurden zunächst für das Warteschlangennetz aufbereitet.

Anschließend konnten die bis dahin hergeleiteten Berechnungsvorschriften auf das Warteschlangennetz angewendet werden. Es wurde festgestellt, daß sich die Komponenten der PAPER-Architektur auf die erwartete Auftragslast abstimmen lassen und durch den mehrfachen Einsatz einzelner Komponenten die Verarbeitungsgeschwindigkeit gesteigert werden kann. Gleichzeitig wurde aber auch aufgedeckt, daß ein schlechtes Verhältnis von Schalt- und Aktivierungszeit die Möglichkeit der parallelen Ausführung von Transitionen schmälert. In diesem Zusammenhang wurde auf die Bedeutung der Konnektivität und der Granularität des PENCIL-Netzes für die Leistungsfähigkeit der PAPER-Architektur hingewiesen.

Die vorliegende Arbeit hat aufgezeigt, wie es durch den Einsatz von warteschlangentheoretischen Mitteln möglich ist, eine abstrakte Kommunikations-Controller-Architektur hinsichtlich der Leistungsfähigkeit zu bewerten und Hilfestellung für die weitere Entwicklung zu leisten. Die dabei erhaltenen Ergebnisse bestätigen die Entwicklung eines Kommunikations-Controllers auf Basis der PAPER-Architektur. Konkretere Aussagen werden erst dann möglich sein, wenn die PAPER-Architektur in einer Netzwerkumgebung emuliert wird, in der das gesamte Protokoll in PENICL/C implementiert ist und die Kommunikation simuliert werden kann. Einen wichtigen Fortschritt des Forschungsprojekts PIKOM wird die Implementierung der PAPER-Architektur auf einem Transputersystem bringen, die augenblicklich durch [Str94] und [Pet94] realisiert wird.

Literaturverzeichnis

- [Bar80] G. Barberis. *A Useful Tool in the Theory of Priority Queueing*. In *IEEE Trans. Comm., COMM-28(9)*, Seiten 1757 – 1761, 1980.
- [Bau90] B. Baumgarten. *Petri-Netze – Grundlagen und Anwendungen*. BI Wissenschaftsverlag, Mannheim, Wien, Zürich, 1990.
- [BB83] J. P. Buzen und A. B. Bondi. *The Response Times of Priority Classes under Pre-emptive Resume in M/M/m Queues*. In *Operations Research, Volume 31*, Seiten 456 – 465, 1983.
- [BCMP75] F. Baskett, K. M. Chandy, R. R. Muntz und F. Palacios. *Open, Closed and Mixed Networks of Queues with Different Classes of Customers*. In *Journal ACM, Volume 22*, Seiten 248 – 260, 1975.
- [Bol89] G. Bolch. *Leistungsbewertung von Rechensystemen mittels analytischer Warteschlangenmodelle*. B.G.Teubner, Stuttgart, 1989.
- [BOP89] H. J. Burkhardt, P. Ochsenschläger und R. Prinoth. *Product Nets, A Formal Description Technique for Cooperating Systems*. GMD-Studien Nr. 165, GMD Darmstadt, September 1989.
- [Bur59] P. J. Burke. *The Output of a Queueing System*. In *Operations Research, Volume 4*, Seiten 699 – 704, 1959.
- [Bur68] P. J. Burke. *The Output Process of a Stationary M/M/s Queueing System*. In *Ann. Math. Statist., Volume 39*, Seiten 1144 – 1152, 1968.
- [Cha72] K. M. Chandy. *The Analysis and Solutions for General Queueing Networks*. In *Sixth Annual Princeton Conference on Information Science and Systems*, Seiten 224 – 228. Princeton University, 1972.
- [Eng90] W. Engbrocks. *Funktionale Simulation der MDMA-Architektur*. Diplomarbeit, RWTH-Aachen, 1990.
- [Eng92] C. Engel. *PENCIL/C, A Language for Concurrent Programming of High Speed Communication Protocols*. Aachener Informatik Berichte Nr. 10/92, RWTH-Aachen, 1992.
- [Eng94] C. Engel. *PAPER, A Multiprocessor High Speed Communication Controller based on Petri Nets*. Aachener Informatik Berichte, RWTH-Aachen, vorraussichtlich 1994.
- [Erl09] A. K. Erlang. *The Theory of Probabilities and Telephone Conversations*. In *Nyt Tidsskrift Matematik, Volume 20*, Seiten 33 – 39, 1909.
- [GN67] W. J. Gordon und G. F. Newell. *Closed Queueing Systems with Exponential Servers*. In *Operations Research, Volume 15*, Seiten 245 – 255, 1967.
- [Gro74] D. Gross. *Fundamentals of Queueing Theory*. John Wiley & Sons, New York, 1974.

- [Gro93] R. Grochtmann. *Entwurf eines Compilers zur Parallel-Programmierung von Hochleistungsprotokollen*. Diplomarbeit, RWTH-Aachen, 1993.
- [Her87] U. Herzog. *Tutorium im Rahmen der 4. GI/ITG Fachtagung für Messung, Modellierung und Bewertung von Rechensystemen*. Erlangen, 1987.
- [Jac57] J. R. Jackson. *Networks of Waiting Lines*. In *Operations Research, Volume 5*, Seiten 518 – 521, 1957.
- [Jac63] J. R. Jackson. *Jobshop-Like Queueing Systems*. In *Management Science, Volume 10*, Seiten 131 – 142, 1963.
- [Jai68] N. K. Jaiswal. *Priority Queues*, Jgg. 50 of *Mathematics in Science and Engineering*. Academic Press, New York, 1968.
- [Jai91] R. Jain. *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, New York, 1991.
- [Ken53] D. G. Kendall. *Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Embedded Markov Chains*. In *Ann. Math. Statist., Volume 24*, Seiten 338 – 354, 1953.
- [Kin90] P. J. B. King. *Computer and Communication Systems Performance Modelling*. Prentice Hall International (UK) Ltd, 1990.
- [Kle75] L. Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, New York, 1975.
- [Kle76] L. Kleinrock. *Queueing Systems, Volume 2: Computer Applications*. John Wiley & Sons, New York, 1976.
- [Kob78] H. Kobayashi. *Modeling and Analysis*. The Systems Programming Series. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1978.
- [Lan92] H. Langendörfer. *Leistungsanalyse von Rechensystemen – Messen, Modellieren, Simulation*. Carl Hanser Verlag, München, 1992.
- [Laz84] E. D. Lazowska. *Quantitative System Performance – Computer System Analysis using Queueing Network Models*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.
- [Lit61] J. D. C. Little. *A Proof for the Queueing Formula $L = \lambda \cdot W$* . In *Operations Research, Volume 9*, Seiten 383 – 387, 1961.
- [Mar90] M. A. Marsan. *Performance Models of Multiprocessor Systems*. MIT Press Series in Computer Systems. The MIT Press, Cambridge, Massachusetts, 3. Auflage, 1990.
- [Pag86] A. Pagnoni. *Petri Nets: Central Models and Their Properties*, Kapitel *Stochastic Nets And Performance Evaluation*. Lecture Notes in Computer Science, Band 254. Springer Verlag, Berlin, 1986.
- [Pet81] J. L. Peterson. *Petri Net Theory and the Modeling of Systems*. Prentice Hall, INC., Englewood Cliffs, N.J. 07632, 1981.

- [Pet94] R. Peters. *Entwurf und Implementierung der Prozeßverwaltung einer petrinetzba-sierten parallelen Kommunikations-Controller-Architektur auf einem Transputersy-tem mit verteiltem Speicher*. Diplomarbeit, RWTH-Aachen, Fertigstellung voraus-sichtlich im Sommer 1994.
- [Rei82] W. Reisig. *Petrinetze, Eine Einführung*. Springer Verlag, 1982.
- [Rup87] M. Rupprecht. *Spezifikation und Bewertung einer Rechnerarchitektur zur Proto-kollimplementierung auf der Basis von Stellen-/Transitionsnetzen*. Diplomarbeit, RWTH-Aachen, 1987.
- [Rup91] M. Rupprecht. *Implementierung und parallele Verarbeitung von Kommunikations-software*. Dissertation, RWTH-Aachen, 1991.
- [Son91] M. Sonnenschein. *Petri-Netze: Eine einführende Übersicht für Studierende der Informatik*. Schriften zur Informatik und angewandten Mathematik, Nr. 153, RWTH-Aachen, Februar 1991.
- [Spa92] O. Spaniol. *Modellierung und Bewertung von Rechensystemen*. Lehrstuhl für Infor-matik IV der RWTH Aachen, Skript zur Vorlesung, WS 91/92.
- [Str94] G. Strauch. *Entwurf und Implementierung von Speichermodellen in einer petri-netzbasierten Controller-Architektur auf einem Transputersystem*. Diplomarbeit, RWTH-Aachen, Fertigstellung voraussichtlich im Sommer 1994.
- [XTP92] *XTP Protocol Definition Rev. 3.6*. Protocol Engines PEI 92-10, 11.01.1992.

A. Die Leistungsgrößen von Wartesystemen

Leistungsgröße		Berechnung
Zustandswahrscheinlichkeit	p_n	$(1 \Leftrightarrow \rho) \cdot \rho^n$
mittlere Verweilzeit	$E[r]$	$\frac{\lambda}{\mu \Leftrightarrow \lambda}$
mittlere Wartezeit	$E[q]$	$\frac{\lambda}{\mu \cdot (\mu \Leftrightarrow \lambda)}$
mittlere Bedienzeit	$E[s]$	$\frac{1}{\mu}$
mittlere Gesamtkundenanzahl	$E[n]$	$\frac{\lambda}{\mu \Leftrightarrow \lambda}$
mittlere Warteschlangenlänge	$E[n_q]$	$\frac{\lambda^2}{\mu \cdot (\mu \Leftrightarrow \lambda)}$
mittlere Kundenanzahl in Bedienung	$E[n_s]$	ρ
Wahrscheinlichkeit, daß sich <i>mindestens</i> k Kunden im System befinden	$P\{n \geq k\}$	ρ^k
Wahrscheinlichkeit, daß sich k Kunden in der Warteschlange befinden	$P\{n_q = k\}$	$\begin{cases} 1 \Leftrightarrow \rho^2 & , k = 0 \\ (1 \Leftrightarrow \rho) \cdot \rho^{k+1} & , k > 0 \end{cases}$

Tabelle A.1: Zusammenfassung der Leistungsgrößen für M/M/1-Wartesysteme

Leistungsgröße		Berechnung
Initialwahrscheinlichkeit	p_0	$\left[\sum_{k=0}^{c-1} \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu}\right)^k + \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \left(\frac{c \cdot \mu}{c \cdot \mu \Leftrightarrow \lambda}\right) \right]^{-1}$
Zustandswahrscheinlichkeit	p_i	$\begin{cases} p_0 \cdot \frac{1}{i!} \cdot \left(\frac{\lambda}{\mu}\right)^i & , 1 \leq i < c \\ p_0 \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{c! \cdot c^{i-c}} & , i \geq c \end{cases}$
mittlere Verweilzeit	$E[r]$	$\left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 + \frac{1}{\mu}$
mittlere Wartezeit	$E[q]$	$\left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0$
mittlere Bedienzeit	$E[s]$	$\frac{1}{\mu}$
mittlere Gesamtkundenanzahl	$E[n]$	$\left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \lambda \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0 + \frac{\lambda}{\mu}$
mittlere Warteschlangenlänge	$E[n_q]$	$\left(\frac{\left(\frac{\lambda}{\mu}\right)^c \cdot \lambda \cdot \mu}{(c \Leftrightarrow 1)! \cdot (c \cdot \mu \Leftrightarrow \lambda)^2} \right) \cdot p_0$
mittlere Kundenanzahl in Bedienung	$E[n_s]$	$\frac{\lambda}{\mu}$
Wahrscheinlichkeit, daß sich <i>mindestens</i> c Kunden im System befinden	$P\{n \geq c\}$	$p_0 \cdot \frac{1}{c!} \cdot \left(\frac{\lambda}{\mu}\right)^c \cdot \left(\frac{c \cdot \mu}{c \cdot \mu \Leftrightarrow \lambda}\right)$

Tabelle A.2: Zusammenfassung der Leistungsgrößen für M/M/c-Wartesysteme

B. Die Ankunftsraten und Verzweigungswahrscheinlichkeiten des Warteschlangennetzes

Ankunftsrate	Berechnung		
	nach Abschnitt 5.3.1	nach Abschnitt 5.3.2	$t_a \stackrel{(5.18)}{=} \delta_f p_{23} \gamma$
λ_1	$\lambda_{05} + \delta_f p_{23} \lambda_{02}$	$\frac{1}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05}$	$\frac{1}{1 \Leftrightarrow t_a} \lambda_{05}$
λ_2	λ_{02}	$\frac{\gamma}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05}$	$\frac{\gamma}{1 \Leftrightarrow t_a} \lambda_{05}$
λ_3	$(1 + \delta_f) \cdot p_{23} \lambda_{02}$	$\frac{(1 + \delta_f) \cdot p_{23} \gamma}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05}$	$\frac{(1 + \delta_f) \cdot p_{23} \gamma}{1 \Leftrightarrow t_a} \lambda_{05}$
λ_4	$\lambda_{05} + 2 \cdot \delta_f p_{23} \lambda_{02}$	$\frac{1 + \delta_f p_{23} \gamma}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05}$	$\frac{1 + t_a}{1 \Leftrightarrow t_a} \lambda_{05}$
λ_5	$(\lambda_{05} + \delta_f p_{23} \lambda_{02}) \cdot (1 + p_{glob})$	$\frac{1 + p_{glob}}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05}$	$\frac{1 + p_{glob}}{1 \Leftrightarrow t_a} \lambda_{05}$
λ_6	$(\lambda_{05} + \delta_f p_{23} \lambda_{02}) \cdot p_{glob}$	$\frac{p_{glob}}{1 \Leftrightarrow \delta_f p_{23} \gamma} \lambda_{05}$	$\frac{p_{glob}}{1 \Leftrightarrow t_a} \lambda_{05}$

Tabelle B.1: Ankunftsraten für die Knoten des Warteschlangennetzes der PAPER-Architektur

Verzweigungswahrscheinlichkeit	Berechnung		
	nach Abschnitt 5.3.1	nach Abschnitt 5.3.2	$t_a \stackrel{(5.18)}{=} \delta_f p_{23} \gamma$
p_{10}	$\frac{\lambda_{05}}{\lambda_{05} + \delta_f p_{23} \lambda_{02}}$	$1 \Leftrightarrow \delta_f p_{23} \gamma$	$1 \Leftrightarrow t_a$
p_{13}	$\frac{\delta_f p_{23} \lambda_{02}}{\lambda_{05} + \delta_f p_{23} \lambda_{02}}$	$\delta_f p_{23} \gamma$	t_a
p_{20}	Ergeben sich aus der Struktur des PENCIL-Netzes		
p_{23}			
p'_{30}	$\frac{\delta_g}{1 + \delta_f}$		
p''_{30}	$\frac{\delta_f}{1 + \delta_f}$		
p_{34}	$\frac{\delta_f}{1 + \delta_f}$		
p_{41}	$\frac{\lambda_{05} + \delta_f p_{23} \lambda_{02}}{\lambda_{05} + 2 \cdot \delta_f p_{23} \lambda_{02}}$	$\frac{1}{1 + \delta_f p_{23} \gamma}$	$\frac{1}{1 + t_a}$
p_{45}	$\frac{\delta_f p_{23} \lambda_{02}}{\lambda_{05} + 2 \cdot \delta_f p_{23} \lambda_{02}}$	$\frac{\delta_f p_{23} \gamma}{1 + \delta_f p_{23} \gamma}$	$\frac{t_a}{1 + t_a}$
p_{54}	$\frac{1}{1 + p_{glob}}$		
p_{56}	$\frac{p_{glob}}{1 + p_{glob}}$		
p_{65}	1		

Tabelle B.2: Verzweigungswahrscheinlichkeiten im Warteschlangennetz der PAPER-Architektur