Approximative Verfahren auf erweiterten Fork/Join–Warteschlangennetzen zur Analyse von Logistiknetzen

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Universität Dortmund
am Fachbereich Informatik

von
Markus Arns

Dortmund
2006

Markus Arns
Lehrstuhl IV - Modellierung und Simulation
Fachbereich Informatik
Universität Dortmund
44227 Dortmund

Tag der mündlichen Prüfung	28.02.2006
Dekan	Prof. Dr. Bernhard Steffen
Gutachter	Prof. Dr.-Ing. Heinz Beilner, Prof. Dr. Peter Buchholz

Danksagung

An dieser Stelle bedanke ich mich bei Herrn Prof. Dr.–Ing. Heinz Beilner für die Betreuung dieser Dissertation und für die Bereitstellung hervorragender Arbeitsbedingungen. Herrn Prof. Dr. Peter Buchholz danke ich für die Übernahme des Zweitgutachtens. Schließlich gilt mein Dank den Mitarbeiterinnen und Mitarbeitern des Lehrstuhls IV „Modellierung und Simulation“ für die freundliche und kooperative Arbeitsatmosphäre.

Inhaltsverzeichnis

I Modellierung und Analyse von Warteschlangennetzen **11**

1 Grundlagen **13**

1.1 Warteschlangennetze – Modellformalismus 14

1.2 Analyseverfahren 16

1.2.1 Analyseverfahren für Produktform–Netze 16

1.2.2 Analyseverfahren für Nicht–Produktform–Netze 18

2 Das Dekompositionsverfahren nach Kühn/Whitt **21**

2.1 Analysierbare Modellklasse 22

2.2 Grundlagen der Analyse 22

2.2.1 Ermittlung des Verkehrsflusses 23

2.2.2 Approximative Analyse von GI/G/1/∞–FCFS Stationen 26

2.2.3 Behandlung direkten Feedbacks 27

2.2.4 Der Dekompositionsalgorithmus 28

3 Erweiterungen des Dekompositionsverfahrens **31**

3.1 Verwendung von Phasenverteilungen 32

3.2 Analyse von PH/PH/1/∞–FCFS Stationen 34

3.2.1 Momente der Populationsverteilung 36

3.2.2 Momente der Durchlaufzeitverteilung 37

3.2.3 Momente des Abgangsprozesses 40
3.3 Analysierbare Modellklasse 41

II Modellierung und Analyse von Fork/Join-Warteschlangennetzen 43

4 Fork/Join–Netze 45
4.1 Das primäre Modell 46

5 Das Upper–Bound Modell 51
5.1 Analyse des Upper–Bound Modells 55
5.1.1 Momente der Kunden–Populationsverteilung 55
5.1.2 Momente der Durchlaufzeitverteilung 56
5.1.3 Der Abgangsprozeß des Upper–Bound Modells 63
5.1.4 Bestimmung der Schranken U 67
5.2 Bewertung des Analyseverfahrens 69
5.2.1 Experimentreihe 1: Homogenes Modell mit zwei parallelen Bedienern . 70
5.2.2 Experimentreihe 2: Inhomogenes Modell mit zwei parallelen Bedienern 76
5.2.3 Experimentreihe 3: Erweitertes inhomogenes Modell 81
5.2.4 Experimentreihe 4: Homogenes Modell mit drei parallelen Bedienern . 83
5.3 Das Upper–Bound Modell in einem Warteschlangenkontext 84
5.3.1 Azyklische Fork/Join–Warteschlangennetze 88
5.3.2 Zyklische Fork/Join–Warteschlangennetze 93

6 Erweiterung auf allgemeine Fork/Join–Netze 97
6.1 $\mathcal{H}/\mathcal{H}/1/\infty$–Aggregate 99
6.2 $\mathcal{H}$/M/1/∞–Aggregate 110
6.3 M/$\mathcal{H}$/1/∞–Aggregate 111
6.4 Ziel–Modellklasse 115

III Anwendungsgebiete **117**

7 Verteilte Computer- und Kommunikationssysteme **119**

7.1 Meta–Suchmaschinen 120

7.1.1 Modellbeschreibung 123

7.1.2 Analyse 124

8 Logistik **131**

8.1 Prozeßketten in der Logistik und das ProC/B–Toolset 132

8.2 Güterverkehrszentrum 135

8.2.1 Modellbeschreibung 136

8.2.2 Analyse 138

8.3 Lieferketten in der Automobilindustrie 142

8.3.1 Modellbeschreibung 143

8.3.2 Analyse 149

9 Zusammenfassung und Ausblick **153**

A Approximation mit Phasenverteilungen **165**

A.1 Halbordnungen auf Verteilungsfunktionen 166

A.2 Monotonieeigenschaften für GI/GI/1–FCFS Systeme 168

A.3 $\mathcal{H}^{1+}$–Verteilungen 170

A.4 $\mathcal{H}^{1-}$–Verteilungen 174

A.5 $\mathcal{H}$–Verteilungen 178

Einleitung

Die rasante technische und technologische Entwicklung im Bereich der Computer–, Kommunikations– und Informationssysteme führt zu einer stetig zunehmenden Präsenz derartiger Systeme in vielen Bereichen des täglichen Lebens. Die zunehmende Verbreitung des Internets mit immer höheren Übertragungsraten und einem immer weiter wachsenden Angebot an Web–Dienstleistungen, die sowohl im privaten als auch im kommerziellen Bereich eine breite Anwendung finden, ist das sicher prominenteste Beispiel für diese Entwicklung. Damit gehen jedoch gleichzeitig stetig steigende Anforderungen hinsichtlich der Leistungsfähigkeit, der Qualität und der Kosten einher, deren optimale Erfüllung letztendlich die Akzeptanz eines Systems bestimmt. Um diesen steigenden Herausforderungen gerecht zu werden, ist die optimale Planung eines Systems in der Entwicklungsphase und die kontinuierliche Überprüfung und Optimierung in der Betriebsphase notwendig. Sowohl in der Planung als auch in der Optimierung von Systemen liegt die übliche Vorgehensweise zunächst in der Definition von Anforderungen an ein zu entwickelndes bzw. bestehendes System und der anschließenden Identifikation und Abbildung relevanter Systemeigenschaften in einem Modell. Spiegelt das Modell das reale System hinsichtlich der gewünschten Informationen ausreichend genau wider, so lassen sich anhand der Analyse des Modells Rückschlüsse auf das reale System ziehen, wie z.B. die Beurteilung der Leistungsfähigkeit und Qualität oder die Erkennung von Schwachstellen und Engpässen.

Die Unterstützung der Modellierung und Analyse technischer Leistungskriterien von Computer– und Kommunikationssystemen ist in der Informatik seit langer Zeit ein wichtiger Themenschwerpunkt und gewinnt aufgrund obiger Schilderungen weiterhin an Aktualität. Die in diesem Kontext betrachteten Systeme lassen sich häufig durch einen diskreten Zustandsraum beschreiben, auf dem in kontinuierlichen zeitlichen Abständen aufgrund des Eintretens diskreter Ereignisse Veränderungen vorgenommen werden. Die wichtigsten Formalismen zur Modellierung und Analyse derartiger *diskreter ereignisgesteuerter dynamischer Systeme* (DEDS)[27] sind Warteschlangennetze, Petri–Netze und die diskrete ereignisgesteuerte Simulation. Die Verfahren unterscheiden sich in dem Detaillierungsgrad der Modelle, in den erzielbaren Analyseresultaten und in der Effizienz der Analyseverfahren hinsichtlich Rechenzeit und Speicherplatzbedarf. Warteschlangennetze zeichnen sich durch die Verfügbarkeit sehr zeit– und platzeffizienter analytisch–algebraischer Analyseverfahren aus, die die exakte Beurteilung quantitativer Leistungskriterien wie Durchsätze, Auslastungen, mittlere Durchlaufzeiten und Populationen erlauben. Sie unterliegen jedoch aufgrund des recht hohen Abstraktionsniveaus gewissen Einschränkungen in der Modellierung und sind damit für Spezialfälle, die diesen Einschränkungen genügen, besonders geeignet. Petri–Netze besitzen einen höheren

Detaillierungsgrad, da sie den Zustandsraum von DEDS sowie die Zustandswechsel direkt abbilden. Die korrespondierenden exakten analytisch–numerischen Analyseverfahren benötigen gegenüber den algebraischen Techniken einen höheren Speicherplatz sowie längere Rechenzeiten, ermöglichen jedoch zusätzlich die Analyse der funktionalen Korrektheit eines Systems. Hinsichtlich der Modellierungsmöglichkeiten ist die Simulation die mächtigste Methode, unterliegt jedoch den beiden anderen Verfahren in dem Sinne, daß die Simulationsergebnisse stets mit einer statistischen Unsicherheit behaftet sind. Zudem lassen sich Modellierungsfehler und funktional unerwünschtes Systemverhalten kaum erkennen. Da Computer– und Kommunikationssysteme häufig den speziellen Anforderungen der Analyseverfahren für Warteschlangennetze oder Petri–Netze genügen, sind beide Methoden zur Modellierung und Analyse von Systemen in diesem Anwendungsbereich besonders gut geeignet.

Aufgrund der Verfügbarkeit dieser Methoden in integrierten Modellierungs– und Analysewerkzeugen [22] und der Interpretation alternativer Anwendungsfälle als diskrete ereignisgesteuerte dynamische Systeme konnten diese Verfahren bereits auf die Planung und Steuerung von Produktionsanlagen übertragen werden. In dem Sonderforschungsbereich 559 „Modellierung großer Netze der Logistik“ (*www.sfb559.uni-dortmund.de*), der im Jahr 1998 an der Universität Dortmund in Kooperation mit dem Fraunhofer Institut für Materialfluß und Logistik eingerichtet wurde, liegt ein Themenschwerpunkt darin, den Nutzen der Analyseverfahren für DEDS im Kontext logistischer Systeme zu evaluieren. Die Logistik nimmt in der modernen Industrielandschaft einen wichtigen Stellenwert ein. So hängt der Erfolg und die Wettbewerbsfähigkeit eines Unternehmens nicht mehr nur von der optimalen Planung, Organisation und Überwachung der eigenen Informationsflüsse und Produktionsprozesse ab, sondern wird in erheblichem Maß von der optimalen Organisation der Schnittstellen zum Kunden und zu Unternehmenspartnern beeinflußt. An diesen Schnittstellen fallen Aufgaben an wie z.B. die Beschaffung, die Distribution, der Warenumschlag und Warentransport, Lagerung und Entsorgung etc. Diese Aufgaben werden von Logistik–Dienstleistern erfüllt und unterliegen gewissen Randbedingungen. So sind Zeitvorgaben und Kostenvorgaben eines Herstellers, vorhandene Transportwege, verfügbare Menge und Art der Transportmittel, begrenztes Personal u.a. zu berücksichtigen. Somit kommt der optimalen Planung, Steuerung und Optimierung logistischer Systeme und damit deren Modellierung und Analyse eine entscheidende Bedeutung zu. Im Rahmen des Sonderforschungsbereichs konnte gezeigt werden, daß die in der Logistik etablierte prozeßorientierte Modellierung (Prozeßketten) prinzipiell mit der Denkweise in diskreten ereignisgesteuerten dynamischen Systemen verträglich ist. Unter diesem Gesichtspunkt ist das ProC/B–Toolset [6] entstanden, in dem die Prozeßketten nach Kuhn [58] derart konkretisiert und formalisiert wurden, daß sie der Modellwelt diskreter ereignisorientierter dynamischer Systeme zugänglich sind. Andererseits wurden jedoch hinsichtlich der Übertragung insbesondere der nicht–simulativen Analyseverfahren für Warteschlangennetze und Petri–Netze gewisse Defizite erkannt, die aus den typischen Systemeigenschaften von Logistiknetzen resultieren. Aufgrund des recht hohen Abstraktionsniveaus eignen sich Warteschlangennetze teilweise gut zur Grobmodellierung von Logistiknetzen. In diesen Fällen liefern die sehr zeiteffizienten Analyseverfahren in kurzer Zeit erste wichtige Analyseresultate. Andererseits bereiten jedoch typische Eigenschaften logistischer Systeme wie die Synchronisation paralleler Abläufe oder die gegenüber Computer– und Kommunikationssystemen im allgemeinen unterschiedliche Charakteristik zeitlicher Abläufe hinsichtlich der Analyseverfahren Schwierigkeiten. Im Bereich der Computer– und Kommunikationssysteme lassen sich zeitliche Abläufe wie z.B. die Ankunftsabstände von Aufträgen an ein System

oder die Ausführungszeiten von Programmen häufig durch negativ–exponentiell verteilte Zufallsvariablen beschreiben. Diese Annahme ist im Bereich Logistik i.a. nicht erfüllt. Im Fall der Petri–Netze erweisen sich die häufig sehr großen Modelle als problematisch, da die zugehörigen Analyseverfahren einen sehr hohen Speicherplatzbedarf haben. Zudem sind meist lediglich negativ–exponentiell verteilte Feuerungszeiten erlaubt. Die Simulation wird in der Logistik zweifelsfrei am häufigsten eingesetzt. Sie birgt jedoch hinsichtlich der Interpretation der Resultate aufgrund der inhärenten Unsicherheit der Simulationsergebnisse auch die größten Risiken. Die Erkennung dieser Defizite wirft unmittelbar die Fragestellung auf, inwieweit sich insbesondere die nicht–simulativen Analyseverfahren hinsichtlich der typischen Eigenschaften von Logistiknetzen anpassen und erweitern lassen.

Diese Arbeit ist mit der Motivation entstanden, einen Beitrag zur Übertragung der effizienten Analyseverfahren für Warteschlangennetze auf den Logistikbereich zu leisten. Das Ziel liegt speziell in der Modellierung der Synchronisation von parallelen Abläufen durch sog. erweiterte Fork/Join–Warteschlangennetze und der Erarbeitung einer geeigneten Analysemethode. Die Motivation zu dieser Arbeit wird weiterhin durch die Relevanz von Fork/Join–Warteschlangennetzen in weiteren wichtigen Anwendungsfeldern bestärkt. So profitiert insbesondere der bereits erwähnte Anwendungsbereich der Computer–, Kommunikations– und Informationssysteme von der Analyse von Fork/Join–Warteschlangennetzen. Beispiele wie Meta–Suchmaschinen, verteilte und parallele Computersysteme, verteilte Datenbanken und RAID–Systeme verdeutlichen die Notwendigkeit dieser Thematik.

Zur Analyse von erweiterten Fork/Join–Warteschlangennetzen bedient sich diese Arbeit eines hybriden dekompositionellen Verfahrens. Dieser Ansatz vereint in gewisser Weise die sehr effizienten analytisch–algebraischen Methoden für Warteschlangennetze mit den zustandsraumbasierten analytisch–numerischen Verfahren. Die Idee des Ansatzes liegt darin, ein Modell in Teilnetze zu zerlegen und diese Teilnetze isoliert voneinander zu analysieren. Die Interaktionen unter den Teilnetzen werden durch Schnittstellenprozesse realisiert, die das Input–/Output–Verhalten der Teilnetze beschreiben. Die lokale Analyse der Teilnetze erfolgt anhand zustandsraumbasierter Verfahren. Dazu wird die Dynamik der Teilnetze, d.h. Input–/Output–Prozesse und Bedienprozesse durch spezielle (unabhängig identisch) phasenverteilte Zufallsvariablen beschrieben bzw. aufgrund der ersten beiden Momente approximiert. Die Verwendung von Phasenverteilungen hat den Vorteil, daß die zustandsraumbasierte Analyse der Teilnetze in eine spezielle Klasse mathematisch sehr gut beherrschbarer Markovketten, sog. Quasi–Birth–and–Death Prozesse (QBD), fällt. QBDs weisen eine regelmäßige Struktur auf, die in der Analyse durch Matrix–geometrischer Verfahren ausgenutzt wird. Dies erlaubt die Behandlung einer Modellklasse, die gegenüber der mit klassischen Verfahren für Warteschlangennetze behandelbaren Klasse deutlich größer ist. Insbesondere wird die Synchronisation paralleler Abläufe beherrschbar. Die Komposition der Einzelresultate zu einem Gesamtergebnis erfolgt gemäß der Prinzipien der analytisch–algebraischen Analyseverfahren. Dabei ist die Annahme die, daß sämtliche Abhängigkeiten unter den Teilnetzen in den Schnittstellenprozessen vollständig erfaßt sind. Zur Bestimmung der Analyseresultate für das Gesamtnetz wird folglich von der Unabhängigkeit der Einzelresultate ausgegangen. Da die Unabhängigkeit i.a. jedoch nicht gegeben ist und auch die Beschreibung der Dynamik durch die speziell verwendeten unabhängig identisch phasenverteilten Zufallsvariablen nur approximativ gilt, liefert das dekompositionelle Verfahren entsprechend approximative Analyseresultate, deren Güte zu bewerten sein wird. In dieser Arbeit wird experimentell belegt, daß in zyklusfreien

Netzen, wie sie typischerweise in der Logistik vorkommen, in vielen Fällen mit einer recht hohen Approximationsgüte zu rechnen ist. In zyklischen Netzen wirkt sich jedoch insbesondere die fälschliche Annahme der Unabhängigkeit der Einzelresultate negativ auf die Qualität der Gesamtergebnisse aus.

Die approximative zustandsraumbasierte Analyse einer recht einfachen Klasse isolierter Fork/Join–Teilnetze wird in [11] vorgestellt. In der vorliegenden Arbeit wird diese Technik in das Dekompositionsverfahren integriert und hinsichtlich der Approximationsgüte bewertet. Diese recht einfachen Strukturen sind jedoch für praktisch relevante Modelle häufig zu restriktiv. Ein weiterer essentieller Beitrag dieser Arbeit ist daher die Erweiterung der einfachen Fork/Join–Struktur auf allgemeinere Fälle. Dies wird durch die Bildung spezieller Aggregate erreicht. Die in der vorliegenden Arbeit entwickelten Aggregate sind dazu geeignet, komplexe Netze durch sehr einfache aus Warteschlangennetzen bekannte Stationstypen derart zu ersetzen, daß unter gleichen Input–Prozessen gleiche Output–Prozesse resultieren. Die nutzbringende Anwendung dieser Aggregattypen im Kontext erweiterter Fork/Join–Warteschlangennetze wird anhand einiger Beispiele aus dem Bereich Logistik und dem Bereich Web–Dienstleistungen belegt.

Zusammenfassend werden in dieser Arbeit die nachfolgend beschriebenen Ziele verfolgt und erreicht. Zunächst wird das Analyseverfahren für Fork/Join–Teilnetze in das Dekompositionsverfahren integriert. Dies ermöglicht die Analyse von einfachen Fork/Join–Warteschlangennetzen. Ferner gelingt es durch die Erarbeitung einer speziellen Aggregierungstechnik, dieses Verfahren um komplexe Fork/Join–Strukturen anzureichern. Die daraus resultierende Analysemethode für erweiterte Fork/Join–Warteschlangennetze erlaubt die Betrachtung einer deutlich größeren Klasse praxisrelevanter Modelle. Damit wird erreicht, daß sich die effizienten Analyseverfahren für Warteschlangennetze, die sich im Bereich der Computer– und Kommunikationssysteme bewährt haben, auf weitere Anwendungsdisziplinen, insbesondere auf den Bereich Logistik übertragen bzw. anpassen lassen. In der Logistik sind komplexe nebenläufige Prozesse inhärenter Bestandteil zahlreicher Anwendungsfälle. Hinsichtlich dieses Anwendungsbereichs liegt der Beitrag der vorliegenden Arbeit damit in der Bereiststellung einer Analysemethode, die sich speziell in der Grobplanung und Optimierung von Logistiknetzen eignet und häufig eine Alternative zu der vielfach in diesem Bereich angewandten Simulation mit ihren oben skizzierten Nachteilen darstellt.

Zur Verfolgung dieser Ziele ist die Arbeit in drei Teile gegliedert. Der erste Teil gibt einen kurzen Überblick über die Modellwelt der Warteschlangennetze und stellt speziell das verwendete Dekompositionsverfahren vor. Der zweite Teil erläutert die Analyse von Fork/Join–Teilnetzen und entwickelt zudem die bereits erwähnten Aggregate. Schließlich präsentiert der dritte Teil einige Beispiele, die den Nutzen des erarbeiteten Analyseverfahrens für erweiterte Fork/Join–Warteschlangennetze demonstrieren.

In Kapitel 1 werden die Grundlagen der Modellwelt Warteschlangennetze skizziert. Es werden die wesentlichen Modellierungsaspekte erläutert und entsprechende Analyseverfahren vorgestellt. Der Fokus liegt dabei insbesondere auf Analyseverfahren, die sich für Fork/Join–Warteschlangennetze eignen bzw. geeignet anpassen lassen. Weiter stellt Kapitel 2 das Dekompositionsverfahren nach Kühn/Whitt vor, das dieser Arbeit als Grundlage zur Analyse von erweiterten Fork/Join-Warteschlangennetzen dienen wird. Es werden die grundlegende Idee erläutert und die wesentlichen Operationen und Analyseschritte beschrieben. Dieses Ver-

fahren wurde bereits vielfach angewendet und hinsichtlich unterschiedlicher Anforderungen angepaßt. Insbesondere wird vielfach die Verwendung von Phasenverteilungen zur Approximation des dynamischen Modellverhaltens eingesetzt. Die Verwendung von Phasenverteilungen im Kontext des Dekompositionsverfahren stellt Kapitel 3 dar. Zudem wird die daraus resultierende Abbildung des dynamischen Verhaltens einer einfachen Warteschlangenstation auf einen QBD und dessen Analyse anhand Matrix–geometrischer Methoden erörtert.

Der Kern der Arbeit beginnt in Kapitel 4 mit der Beschreibung und Modellierung einer einfachen Fork/Join–Station. Es werden spezielle Einschränkungen aufgezeigt, die zur Abbildung dieses Modells auf einen QBD nötig sind. Zudem wird die Analyse des Modells aufgezeigt und in das Dekompositionsverfahren integriert. Schließlich erfolgt eine Bewertung der Analyseresultate im Kontext von Fork/Join–Warteschlangennetzen. Die Erweiterung dieses einfachen Modells auf komplexere Netzstrukturen wird in Kapitel 6 vorgestellt. Dazu werden insbesondere die bereits angesprochenen Aggregate erarbeitet. Im allgemeinen erfolgt die Aggregatberechnung in einem iterativen Verfahren. In zwei Spezialfällen, die ebenfalls aufgezeigt werden, lassen sich die Aggregate jedoch algebraisch ermitteln.

Die Anwendungsbeispiele im dritten Teil der Arbeit werden die entwickelte Technik im Kontext von Web–Dienstleistungen und im Bereich der Logistik erproben. In einem ersten Beispiel wird in Kapitel 7 die Planung einer Meta–Suchmaschine beschrieben. Ferner wird in Kapitel 8 die vorgestellte Technik im Rahmen der Planung von Logistiknetzen angewendet. Dazu werden Lieferketten in der Automobilindustrie sowie wie ein spezieller logistischer Knotenpunkt (Güterverkehrszentrum) betrachtet.

Teil I

Modellierung und Analyse von Warteschlangennetzen

Kapitel 1

Grundlagen

Warteschlangennetze beschreiben Systeme, die eine Menge an Ressourcen umfassen, die wiederum von Nutzern des Systems im Hinblick auf die Erfüllung ihrer Ziele/Wünsche nach einem bestimmten Muster in Anspruch genommen werden. Dabei besteht das Interesse nicht in der Charakterisierung des Verhaltens einzelner, konkret identifizierbarer Nutzer, sondern vielmehr in der Beobachtung typischer, in gewissem Sinne vergleichbarer Nutzer, die ein System in zeitlich wiederkehrenden Abständen nach dem gleichen Muster belasten. Aufgrund der im allgemeinen nur in begrenzter Anzahl zur Verfügung stehenden Ressourcen entstehen folglich immer wieder Wartesituationen. Klassische Beispiele für diese Systeme sind Einkaufsmärkte, in denen Kunden von Mitarbeitern bedient werden, Computersysteme, in denen Jobs nach einem gewissen Schema CPU– und I/O–Anfragen stellen oder auch Produktionssysteme, in denen Rohmaterialien in bestimmter Abfolge von Maschinen zu einem fertigen Produkt bearbeitet werden.

Aus formaler Sicht lassen sich derartige Systeme durch eine Menge von Zuständen beschreiben. Dabei erfaßt ein Zustand die Anzahl der im System anwesenden Nutzer zu einem bestimmten Zeitpunkt bzw. die Anzahl der an jeder einzelnen Ressource wartenden Nutzer zu einem bestimmten Zeitpunkt. Die Systeme sind dynamisch in dem Sinne, daß sie ihre Zustände in kontinuierlichen Zeitabständen durch das Eintreten diskreter Ereignisse verändern. Die diskreten Ereignisse sind z.B. das Eintreffen eines neuen Nutzers im System sowie das Verlassen des Systems und die Belegung sowie Freigabe einer Ressource. Systeme mit diesen Eigenschaften werden mit dem Terminus *Discrete Event Dynamic Systems* (DEDS) [27] bezeichnet.

Die Modellierung mit Warteschlangennetzen hat letztendlich das Ziel, Systeme hinsichtlich ihrer technischen Leistungsfähigkeit zu analysieren. Dazu werden die transiente und die stationäre Analyse unterschieden. Die transiente Analyse beantwortet Fragestellungen hinsichtlich der Startphase eines Systems. In dieser Phase hängt der Systemzustand stark von der Lebensdauer eines Systems ab. Eine interessierende Fragestellung, die die transiente Analyse beantwortet, ist daher z.B. die Bestimmung des Zustands nach t Zeiteinheiten seit Beginn des Systembetriebs. Der Einfluß des Startzustands wird jedoch (unter bestimmten Umständen) mit der Zeit vernachlässigbar sein. Konkret wird sich ab einem gewissen Zeitpunkt die Verteilung der Systemszustände nicht mehr verändern. In diesem Fall hat das Systems einen

stationären Zustand bzw. die stationäre Phase erreicht.

Typische Fragestellungen hinsichtlich der Leistungsfähigkeit eines Systems in der stationären Phase sind:

- Ist das System in der Lage, die aufgetragene Last zu bewältigen?
- Wie lange hält sich ein Nutzer im Mittel in einem System auf?
- Wie lange muß ein Nutzer durchschnittlich auf eine gewisse Ressource warten, bevor er diese selbst beanspruchen kann?
- Wie lang ist durchschnittlich die Warteschlange vor einer bestimmten Ressource?
- Wie hoch sind die einzelnen Ressourcen des Systems ausgelastet?
- An welcher Ressource liegt ein Systemengpaß (Bottleneck) vor?
- Wie wirkt sich die Hinzunahme einer weiteren Ressource eines bestimmten Typs auf die Aufenthaltsdauer der Nutzer im System aus?
- Reichen auch weniger Ressourcen eines bestimmten Typs aus, um eine zufriedenstellende Systemleistung zu erzielen?

Im weiteren Verlauf dieser Arbeit wird ausschließlich die stationäre Analyse von Systemen betrachtet. Zu diesem Zweck wird im folgenden Abschnitt zunächst die bereits skizzierte Modellwelt der Warteschlangennetze präzisiert. Insbesondere wird dazu die genaue Ausprägung der Ressourcen konkreter erläutert sowie auf die Art und Weise ihrer Benutzung eingegangen. Anschließend gibt der Abschnitt 1.2 einen skizzenhaften Überblick über verfügbare Analyseverfahren für Warteschlangennetze. Dazu werden speziell Analyseverfahren betrachtet, die sich im Kontext von Fork/Join–Warteschlangennetzen eignen. Für diese Arbeit ist das in Kapitel 2 vorgestellte Dekompositionsverfahren von zentraler Bedeutung. Diese Methode wird sich unter den skizzierten Analyseverfahren als besonders flexibel erweisen und läßt sich insbesondere für die Analyse von Fork/Join–Warteschlangennetzen anpassen. Weiterhin werden in Kapitel 3 einige Erweiterungen des Dekompositionsverfahren insbesondere hinsichtlich der Verwendung von Phasenverteilungen vorgestellt, da diese in der vorliegenden Arbeit ausgenutzt werden.

1.1 Warteschlangennetze – Modellformalismus

Die Ursprünge der Warteschlangennetze liegen in der Fertigungsplanung (job–shop) im Bereich Operations Research. Ferner haben Warteschlangennetze große Bedeutung in der Modellierung und Leistungsbewertung von Computer– und Kommunikationssystemen erlangt. Ihre hohe Popularität erwächst einerseits aus dem sehr intuitiven Modellierungsansatz und andererseits aus der Existenz sehr effizienter Analyseverfahren, die unter gewissen vereinfachenden Annahmen anwendbar sind. Aufgrund der recht langen Historie dieser Thematik

ist die verfügbare Literatur sehr zahlreich und ausführlich [37, 49, 62, 33, 20]. In den nachfolgenden Ausführungen wird zunächst der Warteschlangenformalismus kurz reflektiert und anschließend auf zugehörige Analyseverfahren eingegangen. Die Darstellung ist an Beilner [20] angelehnt.

Ein Warteschlangennetz läßt sich durch die Beschreibung der statischen Struktur (Maschinenkomponente) und die Spezifikation der dynamischen Struktur (Lastkomponente) charakterisieren. Aus Sicht der Maschinenkomponente besteht ein Warteschlangennetz aus einem Netz aus M Bedieneinrichtungen (Stationen), die willkürlich durchnumeriert und über Pfade untereinander erreichbar sind. Die Stationen ihrerseits besitzen einen Warteraum mit endlicher oder auch unendlicher Kapazität sowie einen oder mehrere gleichartige Bediener. Die Bediener sind durch eine Bediendisziplin charakterisiert, die die Reihenfolge festlegt, in der eintreffende Kunden bedient werden. Sind sämtliche Bediener beschäftigt, so müssen Kunden ggf. in dem Warteraum warten. Ferner legt die Bediengeschwindigkeit die Anzahl der Arbeitseinheiten je Zeiteinheit fest, die ein Bediener zu leisten vermag.

Die Belastung eines Warteschlangennetzes (Lastkomponente) erfolgt durch Kunden, die sich gemäß eines bestimmten Musters in dem Netz bewegen und an die besuchten Stationen Bedienwünsche richten. Die Bedienwünsche der Kunden sind durch kontinuierliche Zufallsvariablen beschrieben, die die zeitliche Inanspruchnahme der Bediener durch einen Kunden angeben. Im allgemeinen besitzen verschiedene Kundentypen an den Stationen eines Netzes unterschiedliche Bedienwünsche, die durch unterschiedliche Zufallsvariablen beschrieben sind. Verschiedene Kundentypen einer Station werden mit dem Begriff Kundenklasse bezeichnet. Die Bewegungsmuster der Kunden werden durch sog. Routing–Wahrscheinlichkeiten $h(i, k; j, l)$ bestimmt, die die Wahrscheinlichkeit angeben, mit der ein Klasse-k Kunde an der Station i nach Beendigung seiner Bedienung der Klasse l an der Station j angehört. Disjunkte Bewegungsmuster, die an allen Stationen disjunkte Klassen einnehmen, werden als Kette bezeichnet. Abschließend bleibt zu klären, auf welche Weise Kunden in ein Warteschlangennetz gelangen. Dazu werden offene und geschlossene Ketten unterschieden. In offenen Ketten treffen Kunden aus der Umgebung an gewissen Stationen des Netzes ein und verlassen das Netz nach Erfüllung ihrer Bedienwünsche. Das Eintreffen der Kunden an einer Station wird ebenfalls durch eine kontinuierliche Zufallsvariable ausgedrückt, die die zeitlichen Abstände aufeinanderfolgender Kunden angibt. Häufig wird der Umgebung eines Netzes die Stationsnummer 0 zugeteilt, so daß das Eintreffen der Kunden an den Stationen und das Verlassen des Netzes in den Routingwahrscheinlichkeiten ausgedrückt werden kann. Im Unterschied zu offenen Ketten existieren in geschlossenen Ketten weder Kundenankünfte noch Kundenabgänge. Stattdessen wird von einer konstant vorhandenen Anzahl von Kunden ausgegangen. Ferner existieren in gemischten Netzen sowohl offene als auch geschlossene Ketten. Die Abbildung 1.1 zeigt beispielhaft ein offenes Warteschlangennetz.

Zur stationären Analyse derartiger Warteschlangennetze hinsichtlich der oben erwähnten Fragestellungen wurden verschiedene Verfahren entwickelt, deren Anwendbarkeit und Effizienz stark von der Ausprägung der statischen und dynamischen Netzstruktur abhängen. Der folgende Abschnitt gibt einen skizzenhaften Überblick über diese Verfahren und fokussiert dabei speziell auf solche Verfahren, die sich im Kontext von Fork/Join–Warteschlangennetzen einsetzen lassen.

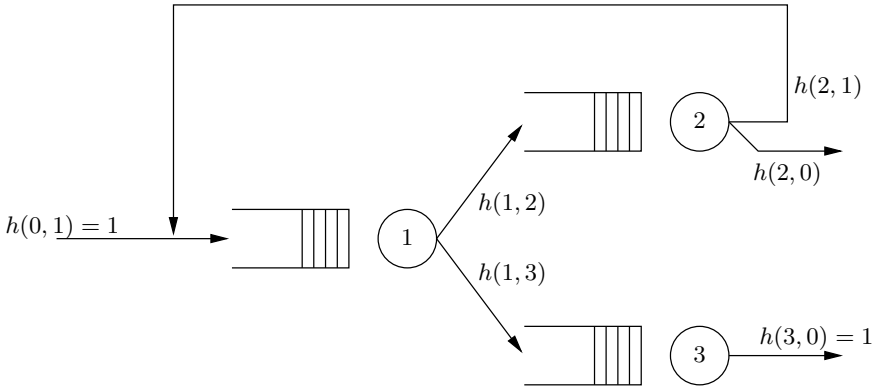

Abbildung 1.1: Beispiel eines Warteschlangenmodells

1.2 Analyseverfahren

Dieser Abschnitt gibt einen Überblick über Analyseverfahren für Warteschlangennetze. Die Verfahren lassen sich grob in exakte Verfahren für Produktform–Netze und approximative Verfahren für Nicht–Produktform–Netze einteilen. Aufgrund der sehr zahlreichen und ausführlichen Literatur zu dieser Thematik wird auf eine detaillierte Beschreibung der Verfahren verzichtet (vgl. hierzu [37, 24, 20, 49, 69, 63, 86, 84, 62, 33, 92]). Aus demselben Grund können die Ausführungen keinen Anspruch auf Vollständigkeit erheben.

1.2.1 Analyseverfahren für Produktform–Netze

Im folgenden werden zunächst die Eigenschaften von Produktform–Netzen aufgezeigt sowie korrespondierende Analyseverfahren vorgestellt. Im Anschluß daran wird ein Überblick über Analyseverfahren für Nicht–Produktform–Netze gegeben, wobei der Fokus insbesondere auf Methoden liegt, die sich zur Analyse von Fork/Join–Warteschlangennetzen eignen.

Produktform–Netze sind die wohl bekannteste Klasse von Warteschlangennetzen. Häufig werden auch die alternativen Bezeichnungen separable Netze oder BCMP–Netze benutzt. Für Produktform–Netze existieren exakte sehr effiziente Analyseverfahren. Die ersten Arbeiten zu diesem Thema wurden von Jackson [44, 45] vorgestellt. Er betrachtete offene Warteschlangennetze mit einer beliebigen Anzahl Stationen. Die nach dem Autor benannten Jackson–Netze besitzen die folgenden Eigenschaften:

- Die Stationen besitzen je eine beliebige feste Anzahl identischer Bediener, die Kunden gemäß der First–Come–First-Serve (FCFS) Bediendisziplin bedienen. Ferner besitzen die Stationen einen unbeschränkten Warteraum.
- Kundenankünfte und Kundenabgänge sind an allen Stationen möglich. Die Kundenankünfte unterliegen einer Poisson–Verteilung.
- Die Bedienzeiten der Kunden an den Stationen sind negativ–exponentiell verteilt.

- Sowohl die Ankunftsraten als auch die Bedienraten an den Stationen können von der Anzahl momentan in der Station anwesender Kunden abhängen (lastabhänge Ankunfts– und Bedienrate).

Jackson zeigte für derartige Warteschlangennetz–Stationen, daß im Falle der Existenz einer stationären Zustandsverteilung diese aus dem Produkt der Zustandsverteilungen der isolierten Stationen resultiert. Dabei repräsentiert ein Zustand des Gesamtnetzes eine bestimmte Anzahl Kunden an den einzelnen Stationen. Dieses Resultat ist recht erstaunlich, da folglich die Berechnung der stationären Zustandsverteilung eines Jackson–Netzes der Rechenregel für die gemeinsame Verteilung unabhängiger Zufallsvariablen gehorcht. Hinsichtlich der Analyse des Gesamtnetzes darf somit die Unabhängigkeit der isolierten Stationen untereinander angenommen werden, was sicher nicht der Fall ist. Diese spezielle Eigenschaft verleiht den Produktform–Netzen ihren Namen.

Gordon/Newell [38] zeigten, daß die Produktform–Eigenschaft auch im Fall geschlossener Warteschlangennetze, die ansonsten dieselben Eigenschaften wie die Jackson–Netze aufweisen, erhalten bleibt. Derartige Produktform–Netze heißen auch Gordon/Newell–Netze.

Eine weitere wichtige Verallgemeinerung, die die Jackson– und Gordon/Newell–Netze einschließt, wurde von Baskett, Chandy, Muntz und Palacios [13] erarbeitet. Sie zeigten, daß die Produktform–Eigenschaft weiterhin für eine erheblich mächtigere Klasse von Warteschlangennetzen gilt. Die nach ihnen benannten BCMP–Netze zeichnen sich durch folgende Eigenschaften aus:

- Netze können gleichzeitig offene und geschlossene Ketten enthalten.
- Es sind die weiteren Bediendisziplinen Processor-Sharing (PS), Infinite–Server (IS) und Last–Come–First–Serve–Preemtive–Resume (LCFS–PR) erlaubt.
- Im Fall der FCFS–Bediendisziplin sind die Bedienzeiten je Station für alle Kundenklassen unabhängig identisch negativ–exponentiell verteilt. Im Fall der übrigen Bediendisziplinen sind beliebige klassenabhängige Bedienzeiten erlaubt.

Grundlage der Analyse offener Produktform–Netze ist die Berechnung der Kundenankunftsprozesse (Verkehrsfluß) an den Stationen. Die Raten der (Poisson'schen) Ankunftsprozesse ergeben sich unter der Kenntnis der Ankunftsraten aus der Umgebung und der Routingwahrscheinlichkeiten aus der Lösung eines linearen Gleichungssystems (Verkehrsflußgleichung). Daraus lassen sich anhand der Bedienraten Leistungsmaße wie Auslastungen, Durchsätze, mittlere Kundenpopulationen und mittlere Durchlaufzeiten sehr leicht für die isolierten Stationen ermitteln.

Die Analyse geschlossener Produktform–Netze gestaltet sich aufwendiger, da aufgrund fehlender Ankunftsprozesse in dem Verkehrsflußgleichungssystem ein Freiheitsgrad übrig bleibt. In der Produktform–Lösung äußert sich dieser Freiheitsgrad in einer Normalisierungskonstanten, die sicherstellt, daß sich die Wahrscheinlichkeiten der (endlich vielen) Zustände des Netzes in der stationären Phase zu 1 summieren. Zur Analyse geschlossener Produktform–Netze wurden zwei zentrale Verfahren vorgestellt.

Der Convolution Algorithm [26] zielt direkt auf die Berechnung der Normalisierungskonstanten ab. Er berechnet die Normalisierungskonstante iterativ über die Anzahl der Stationen eines Netzes und verdankt seinen Namen der Berechnungvorschrift, die der Faltungsoperation von Verteilungsfunktionen ähnlich ist. Die Mean Value Analysis (MVA) [74] vermeidet die explizite Berechnung der Normalisierungskonstanten. Sie basiert auf dem Arrival Theorem [74], das die mittlere Durchlaufzeit eines Kunden an einer Station in Beziehung setzt zu der mittleren Kundenpopulation an derselben Station und einer um 1 geringeren Kundenanzahl im Gesamtnetz. Anhand des Gesetzes von Little können damit iterativ über die Anzahl der Kunden im Netz die oben genannten Leistungsmaße ermittelt werden.

Sowohl von der Mean Value Analysis als auch vom Convolution Algorithm existieren zahlreichen Varianten [31, 29, 77, 28, 1, 34, 98], die sich jeweils für Spezialfälle besonders eignen. Neben diesen exakten Verfahren für Produktform–Netze spielen approximative Verfahren eine wichtige Rolle. Die auf der MVA und dem Convolution Algorithm basierenden approximativen Verfahren dienen einerseits zur Analyse sehr großer geschlossener Produktform–Netze mit vielen Kunden und Kundenklassen, bei denen die exakten Verfahren ineffizient sind oder andererseits zur Analyse von Netzen, die die Produktform–Eigenschaften in einigen Aspekten verletzen. Bekannte approximative Verfahren für Produktform–Netze sind die Bard–Schweitzer Approximation [12, 79] und das SCAT Verfahren [72]. Das DOQ4 Verfahren[21, 22] erlaubt über die Analyse von Produktform–Netzen hinaus zusätzlich die Behandlung von Netzen mit allgemeinen Bedienzeitverteilungen an FCFS–Stationen und Stationen mit Prioritäten.

1.2.2 Analyseverfahren für Nicht–Produktform–Netze

In diesem Abschnitt werden approximative Analyseverfahren für Nicht–Produktform–Netze vorgestellt. Das Repertoire an approximativen Analyseverfahren ist enorm groß. Es umfaßt Methoden für Netze mit speziellen Strukturen, allgemeinen Ankunfts– und Bedienzeitverteilungen, speziellen Bediendisziplinen, Batch–Ankünften und Batch–Bedienungen u.v.a. Im folgenden liegt das Augenmerk auf Verfahren, die sich hinsichtlich der Analyse der in dieser Arbeit betrachteten erweiterten Fork/Join–Warteschlangennetze eignen. Dazu sei zunächst das Modell einer Fork/Join–Station in der Abbildung 1.2 betrachtet.

Die Station enthalte N parallele Stränge, deren konkrete Ausprägung zunächst nicht weiter betrachtet wird. Unmittelbar nach der Ankunft eines Kunden/Auftrags in der Fork/Join–Station werden N Teilaufträge erzeugt, die von je einem der parallelen Stränge bearbeitet werden. Nach der Bearbeitung sämtlicher Teilaufträge verläßt der Kunde/Auftrag die Fork/Join–Station ohne Zeitverzug.

Varki [89, 87, 88] stellt eine Erweiterung der Mean–Value–Analysis (MVA) für geschlossene Produktform–Netze auf geschlossene Fork/Join–Warteschlangennetze mit einer Kundenklasse vor. Die betrachteten Fork/Join–Stationen synchronisieren eine beliebige Anzahl paralleler FCFS–Stationen mit unbegrenztem Warteraum. Die Bedienzeiten sind je Station unabhängig identisch negativ–exponentiell verteilt. Wie bereits im Fall der Mean Value Analysis für Produktform–Netze erwähnt, basiert das wesentliche Konzept auf dem Arrival Theorem, das die mittlere Durchlaufzeit eines Kunden an einer Station zu der mittleren Anzahl der Kunden in der Station bei einer um 1 geringere Kundenanzahl im gesamten Netz setzt. Varki

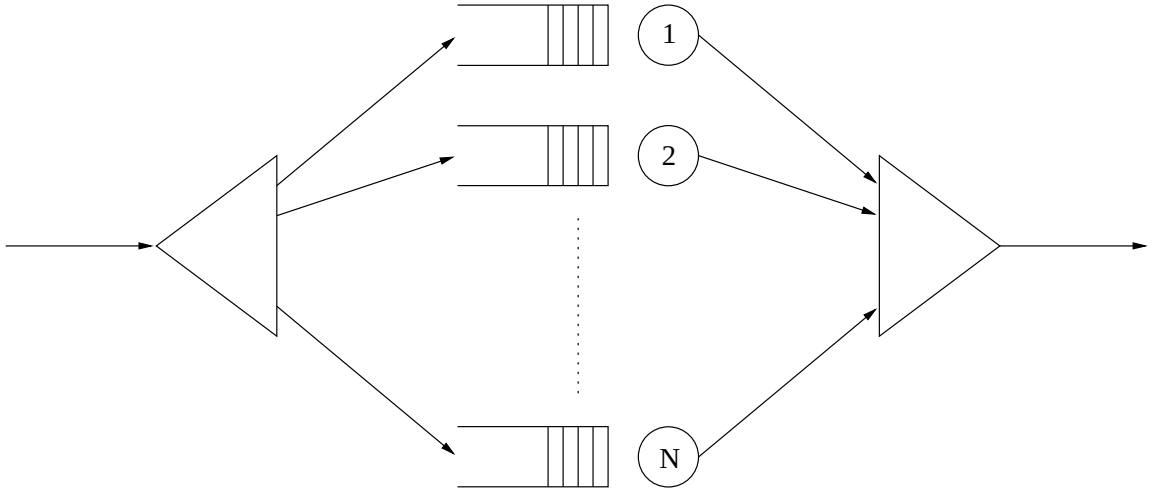

Abbildung 1.2: Modell einer Fork/Join–Station

leitet eine Approximation her, die diese Beziehung auf den betrachteten Typ von Fork/Join–Stationen ausweitet. Für die um diese Approximation angereicherte Mean–Value–Analysis weist Varki experimentell sehr geringe Approximationsfehler nach. Anwendung findet diese Methode insbesondere bei der Leistungsanalyse von RAID–Systemen [90].

Eine Analysemethode für allgemeine geschlossene Nicht–Produktform–Netze basiert auf Arbeiten von Marie [65, 66]. Die Idee des Verfahrens liegt in der Approximation eines allgemeinen geschlossenen Warteschlangennetzes durch ein geschlossenes Produktform–Netz. Dabei wird jede Station des allgemeinen Netzes auf eine FCFS–Station mit lastabhängigen Ankunftsraten sowie lastabhängigen Bedienraten im korrespondierenden Produktform–Netz abgebildet. Ausgehend von einer initialen Belegung der lastabhängigen Bedienraten der Stationen (Aggregate) lassen sich z.B. anhand des Convolution Algorithm die lastabhängigen Ankunftsraten des Produktform–Netzes berechnen. Anhand dieser Raten werden die korrespondierenden Stationen des allgemeinen Warteschlangennetzes isoliert voneinander mit geeigneten Verfahren analysiert. Insbesondere werden die stationären Zustandsverteilungen berechnet. Aus den stationären Zustandsverteilungen der isolierten Stationen lassen sich wiederum neue lastabhängige Bedienraten des Produktform–Netzes ermitteln. Durch wiederholte Durchführung dieser Schritte wird also das Produktform–Netz an das allgemeine geschlossene Warteschlangennetz angepaßt. Die wesentliche Aufgabe hinsichtlich der Analyse allgemeiner geschlossener Warteschlangennetze besteht somit in der Analyse der isolierten Stationen. Baynat und Dallery erweiterten diese Methode auf Mehrketten–Netze [17, 18], indem zu jeder Kette genau ein Produktform–Netz korrespondiert. Ferner nutzen sie das Verfahren von Marie zur Analyse von Fork/Join–Warteschlangennetzen [19]. Dazu reicht es gemäß obiger Schilderungen aus, die stationäre Zustandsverteilung der Fork/Join–Station unter lastabhängigen Ankunftsraten zu berechnen. Koukoumialos und Liberopoulos [52] verwenden dieses Verfahren zur Analyse Kanban–gesteuerter Produktionssysteme.

Baccelli, Massay und Towsley [10] stellen eine Methode für offene sog. azyklische Fork/Join–Warteschlangennetze aus FCFS–Bedieneinrichtungen vor. Aufgrund von stochastischen Ordnungsrelationen (vgl. Anhang A.1) ermitteln sie untere und obere Schranken für die Durch-

laufzeiten von Kunden in einem azyklischen Fork/Join–Warteschlangennetz.

Neben diesen Analyseverfahren für Fork/Join–Warteschlangennetze existieren weiterhin Techniken für die Analyse isolierter Fork/Join–Stationen. Flatto und Hahn [36, 35] betrachten den Spezialfall einer Fork/Join–Station aus $K = 2$ parallelen Bedieneinrichtungen mit FCFS–Bedienung, negativ–exponentiell verteilten Bedienzeiten und Poisson'schem Ankunftsstrom. Sie leiten die generierende Funktion der gemeinsamen Verteilung der Warteschlangenlängen der beiden Bedieneinrichtungen her. Nelson und Tantawi [70] errechnen für den weiter eingeschränkten Fall $K = 2$ mit identisch negativ–exponentiell verteilten Bedienzeiten die mittlere Durchlaufzeit. Für den Fall $K \geq 2$ paralleler Stationen weisen sie untere und obere Schranken für die mittlere Durchlaufzeit nach. Diese Schranken haben die Eigenschaft, daß sie mit steigender Anzahl paralleler Stationen um den gleichen Faktor anwachsen. Nelson und Tantawi nutzen diese Eigenschaft aus und leiten daraus eine approximative Formel (scaling approximation) für die mittlere Durchlaufzeit der betrachteten Fork/Join–Station her. Für Fork/Join–Netze aus $K \geq 2$ M/M/m Stationen mit FCFS–Bedienung und negativ–exponentiell verteilten Zwischenankunftszeiten geben Ko und Serfozo [50] eine Approximation für die Verteilung der Durchlaufzeit an. Im Fall allgemeinerer Ankunfts– und Bedienprozesse sind einige Approximationen sowie Schranken insbesondere für die mittlere Durchlaufzeit bekannt [9, 10, 91, 47, 8, 11, 64].

Im Anwendungsgebiet Produktions– und Auftragssteuerung verwenden Krishnamurthy, Suri und Vernon [57, 54, 56, 55] Fork/Join–Warteschlangennetze zur Modellierung und Leistungsanalyse von Kanban–Kontrollmechanismen. Kanban–Kontrollmechanismen regeln die bedarfsgerechte Fertigung in synchronen Produktionssystemen. Dabei werden in vorgelagerten Produktionsbereichen nur die Mengen produziert, die in nachgelagerten Bereichen verbraucht werden (Pull–Prinzip). Die Steuerung dieser Prozesse läßt sich durch Kanbankarten realisieren, die die Steuerungsdaten dezentral transportieren. Krishnamurthy, Suri und Vernon modellieren derartige Kanban–Systeme durch geschlossene Warteschlangennetze, die durch externe Auftragseingänge und Materialflüsse getriggert werden. Die Analyse dieser Warteschlangennetze erfolgt durch ein approximatives Dekompositionsverfahren, bei dem das dynamische Verhalten der isolierten Stationen auf einen Markov–Prozeß abgebildet wird. Im Unterschied zu dem in dieser Arbeit verfolgten Ansatz werden jedoch die Fork– und Join–Operationen, die die Synchronisation der Materialflüsse und Auftragseingänge mit den Kanban–Steuerungsprozessen repräsentieren, separat voneinander betrachtet.

Generell liegt dieser Arbeit dasselbe Dekompositionsverfahren zugrunde. Die von Kühn/Whitt [59, 60, 94, 93] entwickelte Methode dient zur Analyse einer recht allgemeinen Klasse offener Warteschlangennetze. Sie hat gegenüber den meisten alternativen Techniken den Vorteil, daß sie sich insbesondere hinsichtlich der Analyse von Logistiknetzen, die typischerweise offene Netze sind, flexibel erweitern läßt. Ähnlich wie im Falle des Analyseverfahrens von Marie für geschlossene Warteschlangennetze, erlaubt das Dekompositionsverfahren prinzipiell die Analyse offener Warteschlangennetze mit beliebigen Stationstypen, sofern geeignete Verfahren zur isolierten Analyse der Stationen verfügbar sind. Das Dekompositionsverfahren nach Kühn/Whitt wird im folgenden Kapitel 2 detailliert vorgestellt. Zudem werden in Kapitel 3 Erweiterungen der Methode präsentiert, die sich speziell im Hinblick auf die Analyse von Logistiknetzen als geeignet erweisen werden.

Kapitel 2

Das Dekompositionsverfahren nach Kühn/Whitt

Das Dekompositionsverfahren nach Kühn/Whitt ist ein Analyseverfahren für eine allgemeine Klasse offener Warteschlangennetze. Es wurde Mitte der 70er Jahre von P.J. Kühn vorgestellt [59, 60] und diente zunächst der Analyse von Telekommunikationssystemen. Eine entsprechende Software–technische Realisierung wurde zu Beginn der 80er Jahre von W. Whitt mit dem *Queueing Network Analyzer* präsentiert [94, 93]. Das Ziel des Dekompositionsverfahrens liegt in der zeiteffizienten Analyse von offenen Nicht–Produktform–Netzen unter gewissen Approximationsannahmen. Die wesentliche Lösungsidee besteht darin, ein Warteschlangennetz in eine Menge disjunkter Subnetze zu zerlegen und diese isoliert voneinander bzgl. ihres Ankunftsverhaltens zu analysieren. Die Ankunftsprozesse und die aus der Analyse resultierenden Abgangsprozesse bilden somit die Schnittstelle der Subnetze zu ihrer Umgebung.

Prinzipiell bestehen somit die zentralen Schritte des Dekompositionsverfahrens in einer geeigneten Zerlegung eines Warteschlangennetzes in eine Menge disjunkter Subnetze, die sich unter Zeitaspekten effizient analysieren lassen, und in der Bestimmung des Verkehrsflusses innerhalb des Netzes, d.h. in der Bestimmung der Ankunfts– und Abgangsprozesse der Subnetze. Im allgemeinen ist die isolierte Analyse der Subnetze und die Bestimmung des Verkehrsflusses wechselseitig abhängig, da einerseits die Analyse der Subnetze die Kenntnis des Verkehrsflusses voraussetzt und andererseits die Bestimmung des Verkehrsflusses die vorherige Analyse der Subnetze impliziert.

Zur detaillierten Beschreibung des Dekompositionsverfahrens nach Kühn/Whitt wird im folgenden zunächst die dem Verfahren zugängliche Modellklasse spezifiziert. Auf dieser Basis werden im Anschluß die Grundlagen der Analyse erläutert.

2.1 Analysierbare Modellklasse

Das Dekompositionsverfahren nach Kühn/Whitt eignet sich zur Analyse offener Warteschlangennetze mit einer Kundenklasse und einer festen Anzahl an Stationen. Die Stationen besitzen je einen einzigen Bediener, der an der Station eintreffende Kunden gemäß ihrer Ankunftsreihenfolge bedient (FCFS). Weiterhin zeichnen sich die Stationen durch unbegrenzte Warteräume aus. Kundenübergänge sind potentiell zwischen allen Stationen des Netzes möglich. Insbesondere sind Ankünfte aus der Umgebung und Abgänge in die Umgebung an allen Netzstationen erlaubt.

Das dynamische Verhalten der Kunden im Netz ist durch ihre Zwischenankunfts– und Bedienzeitverteilungen an den Stationen des Netzes sowie durch Routingwahrscheinlichkeiten zwischen den Stationen gekennzeichnet. Im Unterschied zu Produktform–Netzen dürfen die Zwischenankunftszeiten und Bedienzeiten beliebige kontinuierliche nicht–negative Wahrscheinlichkeitsverteilungen besitzen. Kundenübergänge zwischen den Stationen erfolgen dagegen ebenso wie bei Produktform–Netzen gemäß Markov'schem Routing, d.h. nach Beendigung der Bedienung an der i–ten Netzstation wechseln Kunden mit der festen Wahrscheinlichkeit q_{ij} zur j–ten Netzstation für eine weitere Bedienung.

2.2 Grundlagen der Analyse

Zur Analyse eines derartigen Warteschlangennetzes sind bisher keine exakten allgemeinen Verfahren bekannt. Schwierigkeiten bereiten insbesondere die allgemeinen Zwischenankunfts– und Bedienzeitverteilungen. In dem Dekompositionsverfahren nach Kühn/Whitt werden daher die Ankunftsprozesse aller Stationen durch stationäre Erneuerungsprozesse approximiert, d.h. es wird davon ausgegangen, daß die Zwischenankunftszeiten an den Stationen unabhängig identisch verteilt sind. Ebenso wird davon ausgegangen, daß die Bedienzeiten aller Kunden je Station unabhängig identisch verteilt sind. Ferner werden bei der Analyse lediglich die ersten beiden Momente dieser Verteilungen in Form des Erwartungswertes bzw. der Rate und des Variationskoeffizienten berücksichtigt.

Ausgangspunkt der Analyse gemäß Dekomposition nach Kühn/Whitt ist somit die Beschreibung eines offenen Warteschlangennetzes anhand der folgenden Eingangsgrößen:

- Eine feste Anzahl N von FCFS–Einzelbedienern mit unbeschränkter Warteschlange. Die Stationen werden willkürlich von 1 bis N durchnumeriert. Die Umgebung des Netzes erhält die zusätzliche Stationsnummer 0.

- Paare $(\lambda_{0,i}, c_{0,i})$, $i = 1, \ldots, N$. $\lambda_{0,i}$ ist die Rate und $c_{0,i}$ der Variationskoeffizient der externen Zwischenankunftszeitverteilung an Station i, d.h. der Zwischenankunftszeitverteilung von Kunden, die aus der Umgebung an Station i eintreffen.

- Paare (μ_i, c_i), $i = 1, \ldots, N$). μ_i bzw. c_i ist die Rate bzw. der Variationskoeffizient der Bedienzeitverteilung der Kunden an Station i.

- Wahrscheinlichkeiten $q_{i,j}$, $0 \leq q_{i,j} \leq 1$, $1 \leq i \leq N$, $0 \leq j \leq N$. $q_{i,j}$ ist die Wahrscheinlichkeit, mit der ein Kunde nach Beendigung seiner Bedienung an Station i zur Station j für eine weitere Bedienung wechselt. ($q_{i,0}$ ist die Wahrscheinlichkeit, mit der ein Kunde nach Beendigung seiner Bedienung an Station i das Netz verläßt).

Wie bereits zuvor angedeutet, erfolgt die Analyse eines solchen Warteschlangennetzes anhand dreier wesentlicher Operationen. Diese sind die Zerlegung des Netzes in isolierte Subnetze, die Ermittlung des Verkehrsflusses und die Analyse der isolierten Subnetze. Da im vorliegenden Fall alle Stationen denselben Typ besitzen, bietet sich eine Zerlegung des Netzes in seine einzelnen Stationen an. Somit bleiben als zentrale Aufgaben die Analyse der FCFS–Einzelbediener und die Bestimmung des Verkehrsflusses übrig. Hinzu kommt ein weiterer Aspekt, der den direkten Kundenübergang innerhalb einer Station betrifft. Dies ist dann der Fall, wenn es Werte $q_{i,i}, 1 \leq i \leq N$ mit $0 < q_{i,i} < 1$ gibt. In diesem Fall sind der Ankunftsprozeß und Abgangsprozeß der Station i korreliert, und somit ist die Annahme unabhängiger Zwischenankunfts– und Abgangszeiten der i–ten Station nicht haltbar.

Im folgenden wird zunächst auf die Berechnung des Verkehrsflusses eingegangen. Anschließend wird die Analyse der isolierten FCFS–Einzelbediener erläutert, und zuletzt wird am Ende dieses Abschnittes die Behandlung stationsinterner Kundenübergänge behandelt.

2.2.1 Ermittlung des Verkehrsflusses

Zur Ermittlung des Verkehrsflusses eines Warteschlangennetzes sind die Raten λ_i und die Variationskoeffizienten $c_{A,i}$ der unabhängig identisch verteilten Zwischenankunftszeiten aller Station $i = 1, \ldots, N$ des Netzes zu bestimmen. Dazu sei zunächst angenommen, daß geeignete Verfahren zur isolierten Analyse der Subnetze bzw. der Stationen des Netzes bekannt sind und sich damit Leistungsmaße wie Durchlaufzeit, Durchsatz u.a. berechnen lassen. Sei also zunächst eine beliebige Station l dieses Netzes betrachtet. Dann setzt sich der Ankunftsprozeß (mit der Rate λ_l und dem Variationskoeffizienten $c_{A,l}$) dieser Station aus der Überlagerung/Superposition der externen Kundenankünfte aus der Umgebung an dieser Station und aus gewissen Anteilen der Abgangsprozesse der Vorgänger der Station l zusammen. Dieses Szenario ist in der linken Hälfte der Abbildung 2.1 dargestellt.

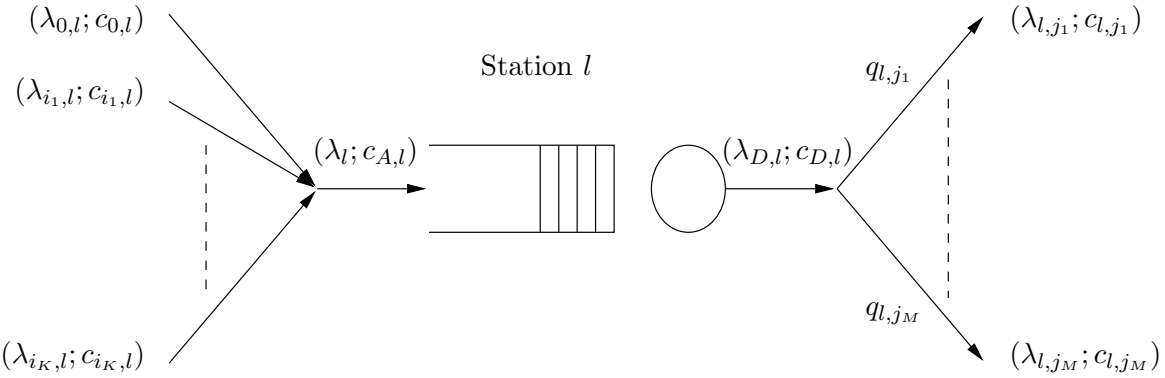

Abbildung 2.1: Superposition und Splitting von Ankunfts- bzw. Abgangsprozessen

Die Ankunftszeiten der Kunden aus der Umgebung an der Station l sind mit der Rate $\lambda_{0,l}$ und dem Variationskoeffizienten $c_{0,l}$ verteilt. Die K ($K \leq N$) Vorgänger der Station l sind genau die Stationen $i_1, \ldots, i_K$, für die $q_{i_k,l} > 0, 1 \leq k \leq K$ gilt. Die Zwischenankunftszeiten von Kunden, die die Station l unmittelbar nach Verlassen der Station i_k besuchen, seien mit der Rate $\lambda_{i_k,l}$ und dem Variationskoeffizienten $c_{i_k,l}$ verteilt (vgl. Abb. 2.1).

Ebenso wie sich der Ankunftsprozeß der Station l durch Superposition mehrerer Teilankunftsprozesse ergibt, resultiert der Einfluß, den der Abgangsprozeß der Station l auf die Ankunftsprozesse der Nachfolger hat, durch Aufspaltung/Splitting des Abgangsprozesses. Diese Situation ist in der rechten Hälfte der Abbildung 2.1 dargestellt. Die Abgangszeiten von Kunden an der Station l seien mit der Rate $\lambda_{D,l}$ und dem Variationskoeffizienten $c_{D,l}$ verteilt. Die M ($M \leq N$) Nachfolger der Station l sind genau die Stationen $j_1, \ldots, j_M$, so daß $q_{l,j_m} > 0, \forall 1 \leq m \leq M$ gilt. Zur Berechnung des Anteils, den der Abgangsprozeß der Station l zum Ankunftsprozeß der Station i_m beträgt, ist der Abgangsprozeß von l gemäß q_{l,j_m} zu splitten. Die Abgangszeiten von Kunden, die unmittelbar nach Verlassen der Station l die Station $j_m, (1 \leq m \leq M)$ besuchen, seien mit der Rate λ_{l,j_m} und dem Variationskoeffizienten c_{l,j_M} verteilt (vgl. Abb. 2.1).

Der folgende Abschnitt geht auf die Bestimmung des Ankunftsprozesses der Station l, charakterisiert durch das Paar $(\lambda_l, c_{A,l})$, und die Bestimmung der (Teil–) Abgangsprozesse, charakterisiert durch die Paare $(\lambda_{l,j_m}, c_{l,j_m}), m = 1, \ldots, M$, ein. Dazu soll nicht die detaillierte Herleitung, sondern die Präsentation der Resultate im Vordergrund stehen. Für detaillierte Informationen sei auf [59] verwiesen.

Bezüglich der Raten erweisen sich sowohl die Superposition als auch das Splitting als recht einfache Operationen. Die Rate λ_l der Superposition ergibt sich durch Summation über alle Raten $\lambda_{i_k,l}, k = 1, \ldots, K$ und $\lambda_{0,l}$, d.h.

$$\lambda_l = \lambda_{0,l} + \sum_{k=1}^{K} \lambda_{i_k,l} \tag{2.1}$$

Die Raten $\lambda_{l,j_m}, m = 1, \ldots, M$ des mit $q_{l,j_m}, m = 1, \ldots, M$ gesplitteten Abgangsprozesses der Station l ergeben sich durch Multiplikation von $\lambda_{D,l}$ mit q_{l,j_M} zu:

$$\lambda_{l,j_m} = q_{l,j_m} \lambda_{D,l} \tag{2.2}$$

Da im vorliegenden Fall ausschließlich Stationstypen betrachtet werden, in denen der Verlust von Aufträgen ausgeschlossen ist, gilt

$$\lambda_l = \lambda_{D,l}.$$

Das Vorgehen zur Bestimmung der Ankunftsraten der Stationen entspricht offenbar dem Vorgehen zur Bestimmung des Verkehrsflusses in Produktform–Netzen. Die Ankunftsraten $\lambda_i, i = 1, \ldots, N$ aller Stationen errechnen sich somit aus der eindeutigen Lösung des folgenden

linearen Gleichungssystems:

$$\lambda_i = \lambda_{0,i} + \sum_{j=1}^{N} q_{j,i}\lambda_j \tag{2.3}$$

Auf ebenfalls recht einfache Weise lassen sich die Variationskoeffizienten $c_{l,j_m}, m = 1, \ldots, M$ aus dem Variationskoeffizienten des Abgangsprozesses der Station l und den Routingwahrscheinlichkeiten q_{l,j_m} ermitteln. So gilt:

$$c^2_{l,j_m} = q_{l,j_m} c^2_{D,l} + (1 - q_{l,j_m}), m = 1, \ldots, M \tag{2.4}$$

Als deutlich aufwendiger erweist sich die Berechnung des Variationskoeffizienten $c_{A,l}$ der Superposition der (Teil–) Ankunftsprozesse. Zu diesem Zweck sei zunächst die Superposition von lediglich zwei Ankunftsprozessen betrachtet. Ferner sei daran erinnert, daß in dem Dekompositionsverfahren nach Kühn/Whitt sämtliche Zwischenankunfts– und Abgangszeiten der Stationen durch stationäre Erneuerungsprozesse approximiert werden. Dann läßt sich die Verteilungsfunktion $F(t)$ der Superposition zweier Ankunftsprozesse mit den Verteilungsfunktionen $F_1(t)$ und $F_2(t)$ und den Ankunftsraten λ_1 und λ_2 angeben zu:

$$F(t) = 1 - \frac{\lambda_1\lambda_2}{\lambda_1 + \lambda_2} \left((1 - F_1(t)) \int_t^\infty (1 - F_2(s))ds + (1 - F_2(t)) \int_t^\infty (1 - F_1(s))ds \right) \tag{2.5}$$

Zum weiteren Vorgehen werden beide Ankunftsprozesse in Abhängigkeit ihres Variationskoeffizienten durch geeignete Phasenverteilungen approximiert, von denen die Verteilungsfunktionen bekannt sind. Auf die Approximation mit Phasenverteilungen wird in Anhang A detailliert eingegangen. Somit läßt sich also die Verteilungsfunktion der Superposition zweier Ankunftsprozesse approximativ ermitteln. Daraus wiederum kann der Variationskoeffizient der Superposition berechnet werden. Die Rate der Superposition ergibt sich aus der Summe der Einzelraten. Die Superposition von mehr als zwei Ankunftsprozessen wird durch Zusammenfassen von jeweils zwei Prozessen und wiederholter Anwendung des Verfahrens bestimmt bzw. approximiert.

Mit diesen Vorarbeiten lassen sich nun die Variationskoeffizienten $c_{A,i}, i = 1, \ldots, N$ aller Stationen des Netzes ermitteln. Im Fall azyklischer Netze werden zunächst die Stationen mit externen Ankunftsprozessen betrachtet. Für diese Stationen lassen sich die ersten beiden Momente der Abgangsprozesse ermitteln, da nach Voraussetzung geeignete Analyseverfahren existieren. Aus diesen Abgangsprozessen werden nacheinander unter Anwendung der Splitting– und Superpositions–Operation die ersten beiden Momente der Ankunfts– und Abgangsprozesse der übrigen Stationen berechnet.

Im Fall zyklischer Netze ist die Berechnung der $c_{A,i}$ nicht auf derart einfache Weise durchführbar. Stattdessen wird ein iteratives Verfahren angewandt, bei dem zunächst alle Variationsko-

effizienten $c_{i,j}$ mit 1 initialisiert werden, sofern ein Kundenübergang von der Station i zur Station j möglich ist. Die Variationskoeffizienten $c_{0,i}$ der externen Zwischenankunftszeiten bleiben unverändert. Damit lassen sich dann in einem ersten Schritt anhand der Superpositions–Operation und der im Vorfeld berechneten Raten $\lambda_{i,j}$ (vgl. Gleichung 2.2) die Variationskoeffizienten $c_{A,i}$ der Zwischenankunftszeitverteilungen aller Stationen berechnen. Im zweiten Schritt werden durch Analyse der isolierten Stationen die Variationskoeffizienten $c_{D,i}$ der Abgangsprozesse aller Stationen ermittelt. Daraus lassen sich mittels der Splitting–Operation neue Variationskoeffizienten $c'_{i,j}$ bestimmen, für die das Iterationsschema mit Schritte 1, der Superpositions–Operation, wiederholt wird, bis die $c_{A,i}$ ein vorgegebenes Konvergenzkriterium erfüllen. Angemerkt sei hierzu jedoch, daß die Konvergenz dieses Iterationsschemas nicht gesichert ist (vgl. [94]).

2.2.2 Approximative Analyse von GI/G/1/∞–FCFS Stationen

Zur Analyse einer GI/G/1/∞–FCFS Station sei also davon ausgegangen, daß sowohl die Rate λ und der Variationskoeffizient c_A der Zwischenankunftszeitverteilung als auch die Rate μ und der Variationskoeffizient c_S der Bedienzeitverteilung bekannt sind. Das Interesse besteht insbesondere in der Bestimmung der mittleren Durchlaufzeit $\overline{V}$ und in der Bestimmung des Abgangsprozesses bzw. der Verteilung D des Durchsatzes bzgl. der ersten beiden Momente. Die exakte Analyse der Leistungsmaße dieses Stationstyps ist lediglich für einige wenige Spezialfälle, wie z.B. negativ–exponentiell verteile Zwischenankunfts– und Bedienprozesse, bekannt. Im Fall beliebig verteilter Zwischenankunfts– und Bedienprozesse sind jedoch keine exakten Aussagen über die interessierenden Größen möglich. Jedoch existieren approximative Formeln, die auf Basis der ersten beiden Momente der Zwischenankunfts– und Bedienzeitverteilung Aussagen über die mittlere Durchlaufzeit und die ersten beiden Momente des Abgangsprozesses treffen.

So läßt sich die mittlere Wartezeit $\overline{W}$ der Kunden vor ihrer Bedienung anhand der Krämer/Langenbach–Belz Formel [53] wie folgt approximieren:

$$\overline{W} = \frac{1}{\mu}\frac{\rho}{2(1-\rho)}(c_A^2 + c_S^2)g(\rho, c_A^2, c_S^2) \tag{2.6}$$

Dabei ist:

$$g(\rho, c_A^2, c_S^2) = \begin{cases} exp\left(-\frac{2(1-\rho)}{3\rho}\frac{(1-c_A^2)^2}{c_A^2+c_S^2}\right) & \text{, falls } c_A^2 < 1 \\ exp\left(-(1-\rho)\frac{c_A^2-1}{c_A^2+c_S^2}\right) & \text{, falls } c_A^2 \geq 1 \end{cases} \tag{2.7}$$

Die mittlere Durchlaufzeit $\overline{V}$ ergibt sich daraus zu

$$\overline{V} = \overline{W} + \frac{1}{\mu}.$$

Ebenso läßt sich der Variationskoeffizient des Abgangsprozesse anhand Marshall's Formel [68]

approximieren. (Die Rate des Abgangsprozesse entspricht offensichtlich der Rate λ des Ankunftsprozesses).

$$c_D^2 \approx c_A^2 + \rho^2(c_S^2 - c_A^2) \tag{2.8}$$

2.2.3 Behandlung direkten Feedbacks

Ein direktes Feedback bedeutet, daß Kunden mehrfach hintereinander dieselbe Station i besuchen, ohne zwischenzeitlich eine andere Station $j \neq i$ zu besuchen. In diesem Fall gilt $q_{ii} > 0$. Die aufeinanderfolgenden Bedienprozesse eines Kunden an solch einer Station i finden i.a. nicht direkt hintereinander statt, sondern werden durch die Bedienung weiterer Kunden unterbrochen.

Das Problem einer Station mit direktem Feedback besteht darin, daß der Abgangsprozeß dieser Station mit dem Ankunftsprozeß der Station korreliert ist. Folglich ist die Annahme der Unabhängigkeit aufeinanderfolgender Zwischenankunftszeiten bzw. Abgangszeiten an dieser Station nicht gerechtfertigt. Dieses Problem wird dadurch behandelt, daß die aufeinanderfolgenden Bedienprozesse eines Kunden an einer Station mit direktem Feedback zu einem einzigen modifizierten Bedienprozeß zusammengefaßt werden. Bei dieser Vorgehensweise bleibt offensichtlich die Tatsache unberücksichtigt, daß zwischen den einzelnen Bedienungen eines Kunden i.a. weitere Kunden bedient werden.

Zur Berechnung des modifizierten Bedienprozesses von Kunden an einer Station i mit direktem Feedback wird ausgenutzt, daß die Anzahl X der aufeinanderfolgenden Besuche geometrisch mit dem Parameter $(1 - q_{i,i})$ verteilt ist. Es gilt also:

$$p_x := P\{X = x\} = \begin{cases} 0 & \text{, falls } x = 0 \\ (1 - q_{i,i})q_{i,i}^{x-1} & \text{, falls } x \geq 1 \end{cases} \tag{2.9}$$

Die generierende Funktion $G_X(z)$ der Verteilung von X ist damit:

$$G_X(z) = \sum_{x=0}^{\infty} p_x z^x = \frac{(1 - q_{i,i})z}{1 - q_{i,i}z} \tag{2.10}$$

Die Laplace–Transformierte $L(s)$ des modifizierten Bedienprozesses ergibt sich aus der Laplace–Transformierten des ursprünglichen Bedienprozesses zu:

$$L(s) = \sum_{x=0}^{\infty} p_x (H(s))^x \tag{2.11}$$

Mit dem Ergebnis 2.10 folgt:

$$L(s) = G_X(H(s)) = \frac{(1 - q_{i,i})H(s)}{1 - q_{i,i}H(s)} \tag{2.12}$$

Daraus ergeben sich die Rate μ_i^* und der Variationskoeffizient $c_{S,i}^*$ des modifizierten Bedienprozesses von Kunden an der Station i zu:

$$\mu_i^* = (1 - q_{i,i})\mu_i \tag{2.13}$$

$$c_{S,i}^* = \sqrt{(1 - q_{i,i})c_{S,i}^2 + q_{i,i}} \tag{2.14}$$

Dementsprechend sind die Routingwahrscheinlichkeiten $q_{i,j}$ der Station i folgendermaßen anzupassen:

$$q_{i,j}^* = \begin{cases} 0 & \text{, falls } i = j \\ \frac{q_{i,j}}{1-q_{i,i}} & \text{, falls } i \neq j \end{cases} \tag{2.15}$$

2.2.4 Der Dekompositionsalgorithmus

Die Resultate der Abschnitte 2.2.1 bis 2.2.3 lassen sich nun in folgendem Algorithmus zusammenfassen:

Algorithmus 2.1 *Dekomposition nach Kühn/Whitt*

1. *Ersetze die Bedienzeitverteilungen aller Stationen mit direktem Feedback durch die entsprechenden modifizierten Bedienzeiten anhand der Gleichungen 2.13 und 2.14, und passe die Routing–Wahrscheinlichkeiten gemäß 2.15 an.*

2. *Berechne den Verkehrsfluß des Warteschlangennetzes bzgl. der Raten anhand des linearen Gleichungssystems 2.3.*

3. *Initialisiere die Variationskoeffizienten der internen Kunden–Übergangsprozesse mit 1, d.h. setze* $c_{i,j} = 1 \Leftrightarrow q_{i,j} > 0 \quad \forall i, j = 1, \ldots, N$.

4. *Berechne die Variationskoeffizienten* $c_{A,i}, i = 1, \ldots, N$ *der Ankunftsprozesse aller Stationen des Netzes anhand der Superpositions–Operation (vgl. Abschnitt 2.2.1).*

5. *Approximiere die Variationkoeffizienten* $c_{D,i}, i = 1, \ldots, N$ *der Abgangsprozesse aller Stationen anhand Marshall's Formel 2.8.*

6. *Bestimme anhand der Splitting–Operation die Variationskoeffizienten $c_{i,j}$ der internen Übergangsprozesse neu, und wiederhole die Schritte 4 bis 6 solange, bis für alle Paare (i, j) mit $q_{i,j} > 0$ die Differenz der Werte $c_{i,j}$ aus zwei aufeinanderfolgenden Durchläufen des Iterationsschemas eine fest gewählte Genauigkeitsschranke unterschreiten.*

7. *Berechne die mittleren Durchlaufzeiten der Kunden an allen Stationen anhand der Krämer/Langenbach–Belz Approximation (vgl. 2.6 und 2.7).*

Kapitel 3

Erweiterungen des Dekompositionsverfahrens

Das Dekompositionsverfahren nach Kühn/Whitt dient dieser Arbeit als Basisverfahren zur Analyse von offenen Nicht–Produktform Warteschlangennetzen. Das Ziel besteht letztendlich darin, die Klasse der analysierbaren Warteschlangennetze auf Fork/Join–Netze zu erweitern. Zu diesem Zweck sind zunächst einige Anpassungen bzw. Erweiterungen des Dekompositionsverfahrens nötig. Zur Verdeutlichung dieser notwendigen Anpassungen werden im folgenden die zentralen Prinzipien des Verfahrens nochmals zusammengefaßt.

Prinzipien des Dekompositionsverfahrens nach Kühn/Whitt:

1. Zerlegung eines Warteschlangennetzes in eine Menge disjunkter Subnetze/Stationen, deren Schnittstellen durch die Ankunfts– und Abgangsprozesse gebildet werden.
2. Approximation aller Zwischenankunfts– und Abgangsprozesse durch stationäre Erneuerungsprozesse.
3. Analyse der isolierten Stationen anhand einfacher approximativer Formeln.
4. Betrachtung der Zwischenankunfts– und Abgangsverteilungen sowie der Bedienzeitverteilungen ausschließlich bzgl. ihrer Raten und Variationskoeffizienten.

Die maßgebliche Arbeit von Kühn [59] bildet in der Folgezeit die Ausgangsbasis zahlreicher Forschungsarbeiten auf dem Gebiet der Analyse von offenen Nicht–Produktform–Warteschlangennetzen durch Dekomposition. Die weiterführenden Arbeiten greifen die oben genannten zentralen Prinzipien auf und verfolgen stets das Ziel, einige der erwähnten Approximationen zu verbessern oder aber die Klasse der mit dem Verfahren analysierbaren Warteschlangennetze auszuweiten.

So betrachtet Whitt [94] Netze mit mehreren Kundenklassen und Stationen mit mehreren identischen Bedienern. Ebenso erlaubt er die Bündelung und Zerteilung von Aufträgen/Kunden

an den Stationen. Die Bündelung bzw. Zerteilung von Aufträgen wird dadurch realisiert, daß die Zwischenankunftsraten $\lambda_i, i = 1, \ldots, N$ der Stationen mit Faktoren γ_i multipliziert werden, für die im Fall einer Bündelung $\gamma_i < 1$ und im Fall einer Zerteilung $\gamma_i > 1$ gilt. Den Erwartungswert $E[W]$ der Verteilung der Wartezeit eines entsprechenden GI/G/m–Systems berechnet Whitt aus dem Erwartungswert eines M/M/m–Systems durch Multiplikation mit dem Faktor $\frac{c_A^2+c_S^2}{2}$. c_A und c_S sind die Variationskoeffizienten des Ankunftsprozesses und des Bedienprozesses der GI/G/m–Station. Den Variationskoeffizienten der Wartezeitverteilung approximiert er durch den Variationskoeffizienten des entsprechenden M/M/m–Systems. Diese Approximation des Erwartungswertes und des Variationskoeffizienten der Wartezeitverteilung ist im Fall hoher Auslastungen und in dem Fall, daß die Variationskoeffizienten c_A und c_S nahe bei 1 liegen, sehr gut. Im allgemeinen weichen die tatsächlichen Werte jedoch erheblich von diesen Approximationen ab. Die Behandlung von Mehrklassen–Warteschlangennetzen führt Whitt auf Einklassen–Netze zurück, indem er die Kunden unterschiedlicher Klassen an einer Station zu einem typischen Kunden aggregiert. Die Berechnung der Bedienzeitverteilung dieses typischen Kunden erfolgt durch Gewichtung der Bedienzeitverteilungen der unterschiedlichen Kundenklassen aufgrund ihrer Ankunftsraten. Die Resultate dieser Approximation sind im allgemeinen ebenfalls recht weit von den tatsächlichen Werten entfernt. Verfeinerungen dieses Ansatzes werden in [96, 23] insbesondere hinsichtlich der Ermittlung Klassen–spezifischer Durchlaufzeiten und Abgangsprozesse vorgestellt.

3.1 Verwendung von Phasenverteilungen

In [40, 41, 76] wird das Dekompositionsverfahren nach Kühn/Whitt derart erweitert, daß die Ankunfts– und Bedienprozesse der FCFS–Stationen durch Phasenverteilungen hinsichtlich der ersten beiden Momente approximiert werden. Die Approximation durch Phasenverteilungen ist durch zwei Aspekte motiviert. Zum einen zeigte Cox [32], daß sich kontinuierliche nicht–negative Verteilungen mit rationaler Laplace–Transformierten, die ausschließlich negative reellwertige Polstellen besitzt, hinsichtlich ihrer Momente beliebig genau durch bestimmte Phasenverteilungen approximieren lassen. Zum zweiten erlaubt diese Darstellung eine besonders kompakte Beschreibung des dynamischen Verhaltens zahlreicher Stationstypen in Form von Quasi–Birth–and–Death Prozessen (QBD) [73]. QBDs lassen sich anhand Matrix–geometrischer Methoden auf numerische Weise exakt analysieren. Auf diese Weise lassen sich im Sinne der Approximation durch Phasenverteilungen die Prinzipien 3 und 2 des Dekompositionsverfahrens durch exakte Ergebnisse ersetzen.

In weiteren Ansätzen [75, 42, 43] wurde die Idee der Approximation durch Phasenverteilungen aufgegriffen und um sog. Markovian Arrival Processes (MAPs) erweitert. Im Gegensatz zu Phasenverteilungen erlauben MAPs die Beschreibung von Abhängigkeiten zwischen aufeinanderfolgenden Ankünften oder Bedienungen an einer Station. Somit läßt sich das Verhalten realer Systeme exakter darstellen, als dies mit Phasenverteilungen möglich ist. In dieser Arbeit wird auf die Verwendung von MAPs aus verschiedenen Gründen verzichtet. Einerseits erhöhen MAPs im Vergleich zu Phasenverteilungen die Komplexität der in den folgenden Abschnitten betrachteten QBDs und liefern hinsichtlich der grundlegend vorgestellten Konzepte keine wesentlichen neuen Erkenntnisse. Sie eignen sich somit besonders zur Verfeinerung der in dieser Arbeit vorgestellten Verfahren. Eine zweite Begründung liegt darin, daß eine exaktere

Repräsentation der Verteilungen von Zwischenankunfts– und Bedienzeiten voraussetzt, daß entsprechendes Wissen hinsichtlich des modellierten realen Systems vorliegt. Dieses Wissen ist jedoch in vielen Anwendungsfällen, gerade im Bereich Logistik, nur unzureichend vorhanden, so daß MAPs die Gegebenheiten eines realen Systems nicht zwingend exakter beschreiben als Phasenverteilungen.

Im folgenden werden die Auswirkungen, die die Verwendung von Phasenverteilungen auf das Dekompositionsverfahren hat, erläutert. Dazu werden die Begriffe Phasenverteilung, Quasi–Birth–and–Death Prozeß sowie Matrix–geometrische Methoden kurz skizziert. Ausführlichere Informationen zu Phasenverteilungen gibt Anhang A. Eine detaillierte Einführung in die Thematik der Quasi–Birth–and–Death Prozesse sowie die korrespondierenden Matrix–geometrischen Methoden geben [73] und [61].

Eine phasenverteilte Zufallsvariable setzt sich aus einer endlichen Anzahl negativ–exponentiell verteilter Zufallsvariablen zusammen und läßt sich somit durch eine endliche zeitkontinuierliche Markovkette beschreiben. Die endliche Markovkette ist durch eine Matrix–Vektor–Darstellung (T, τ) charakterisiert, wobei T die Ratenmatrix der Zustandsübergänge und τ die initiale Verteilung der Zustände der Markovkette repräsentieren. Die Approximation einer kontinuierlichen nicht–negativen Verteilung durch eine Phasenverteilung erfolgt häufig auf der Grundlage des Variationskoeffizienten c der Verteilung. Im Fall $c = 1$ ist offensichtlich die Wahl einer Negativ–Exponentialverteilung (Phasenverteilung mit genau einer Phase) geeignet. Verteilungen mit einem Variationskoeffizienten $c > 1$ lassen sich bzgl. der ersten beiden Momente exakt durch eine Phasenverteilung mit genau zwei alternativen Phasen (Hyper–Exponentialverteilung) darstellen. Hingegen können Verteilungen mit einem Variationskoeffizienten $0 < c < 1$ durch $K = \lceil \frac{1}{c^2} \rceil$ hintereinandergeschaltete negativ–exponentiell verteilte Phasen (Hypo–Exponentialverteilung) beschrieben werden, von denen $K - 1$ identische Raten besitzen. Auf die Approximation mit Phasenverteilungen sowie auf die Hyper–Exponentialverteilung und die Hypo–Exponentialverteilung wird im Anhang A detailliert eingegangen.

Anhand der Approximation der Verteilungen der Zwischenankunfts– und Bedienzeiten durch Phasenverteilungen ist das dynamische Verhalten einer recht mächtigen Klasse von Stationstypen durch QBDs beschreibbar. QBDs sind Markovketten, deren Generatormatrizen eine spezielle Tridiagonalgestalt mit sich wiederholenden Diagonalblöcken besitzen und die besonders angenehme mathematische Eigenschaften aufweisen. Insbesondere läßt sich die stationäre Zustandsverteilung (soweit diese existiert) eines QBDs anhand Matrix–geometrischer Methoden recht leicht ermitteln. Die sich wiederholenden Tridiagonalblöcke der Generatormatrix eines QBDs erlauben eine Darstellung des Vektors der stationären Grenzverteilung auf eine ähnliche Weise, wie dies z.B. für M/M/1–∞ Stationen möglich ist. In M/M/1–∞ Stationen sind die Zustände π_i des stationären Systems geometrisch mit dem Parameter ρ (Auslastung) verteilt, d.h.

$$\pi_0 = 1 - \rho \tag{3.1}$$

$$\pi_i = \pi_0 \rho^i \tag{3.2}$$

Für QBDs läßt sich eine ähnliche Beziehung unter den stationären Zustandswahrscheinlich-

keiten angeben, bei der jedoch an die Stelle der Auslastung ρ eine gewisse Matrix R tritt. Dieser Ähnlichkeitsbeziehung verdanken die Matrix–geometrischen Methoden ihren Namen. Der stationäre Zustandsvektor bildet die Basis zur Ermittlung verschiedener Leistungsmaße einer Station. QBDs sowie deren mathematische Analyse mit Matrix–geometrischen Methoden werden in [73] und [61] ausführlich behandelt.

Die Abbildung des dynamischen Systemverhaltens auf einen QBD sowie dessen exakte Analyse wird nachfolgend anhand der in Abschnitt 2.2.2 approximativ behandelten GI/G/1/∞–FCFS Station illustriert.

3.2 Analyse von PH/PH/1/∞–FCFS Stationen

In diesem Abschnitt wird die Abbildung einer PH/PH/1/∞–FCFS Station auf einen QBD sowie die exakte Analyse des QBDs vorgestellt. Die betrachtete Station besitzt eine unbegrenzte Anzahl an Warteplätzen und einen Bediener, der gemäß Reihenfolge erhaltender Bediendisziplin (FCFS) arbeitet. Die unabhängig identisch verteilten Zwischenankunftszeiten sowie die ebenfalls unabhängig identisch verteilten Bedienzeiten seien durch geeignete Phasenverteilungen mit den Repräsentationen (A, α) und (B, β) beschrieben. Folglich entspricht die Zwischenankunftszeitverteilung (und ebenso die Bedienzeitverteilung) der Verteilung der Absorbtionszeit einer durch die Generatormatrix A und den Startvektor α beschriebenen endlichen absorbierenden Markovkette (vgl. Def. A.1). Die Phasenverteilung (A, α) bzw. (B, β) habe n bzw. m Phasen. Für derartige Stationstypen ist die Notation PH/PH/1/∞–FCFS gebräuchlich.

Das Ziel der Betrachtungen besteht in der stationären Analyse dieser Station. Dabei liegt der Fokus insbesondere auf der Ermittlung der Verteilungscharakteristika der Kundenpopulation, der Durchlaufzeit und des Abgangsprozesses. Ein geeignetes mathematisches Modell, das diesen Stationstyp im Hinblick auf die interessierenden Analyseresultate vollständig abbildet, ist die (unendliche) Markovkette. Ein Zustand dieser Markovkette repräsentiert die Anzahl der momentan in der Station anwesenden Kunden. Der Zustandsraum Z ist damit wie folgt definiert:

$$Z = \{z : z \in I\!N_0\} \tag{3.3}$$

Diese Wahl von Z gewährt einen recht groben Einblick in das System, da insbesondere die aktuellen Phasen des Ankunfts– und des Bedienprozesses verborgen bleiben. Um den Detaillierungsgrad des Zustandsraums deutlich hervorzuheben, werden die Zustände $z \in Z$ mit dem Begriff *Makrozustände* bezeichnet. Im Unterschied dazu werden Zustände, die die aktuellen Phasen des Ankunfts– und Bedienprozesses berücksichtigen, *Mikrozustände* genannt. Aufgrund der Darstellung des Ankunftsprozesses durch n Phasen und des Bedienprozesses durch m Phasen enthält der Makrozustand 0 genau n Mikrozustände und alle übrigen Makrozustände $z \geq 1$ genau $n \cdot m$ Mikrozustände.

Der Detaillierungsgrad von Z spiegelt sich in der Darstellung der Zustandsübergänge $Q[z, z']$ zwischen zwei Makrozuständen $z, z' \in Z$ wieder. Die Verwendung der eckigen Klammern verdeutlicht, daß es sich bei der Notation von $Q[z, z']$ um die Spezifikation einer Matrix handelt. Die Zustandsübergänge zwischen zwei disjunkten Makrozuständen $z \neq z', z, z' \in Z$

sind dementsprechend auf folgende Weise definiert:

$$Q[z,z'] = \begin{cases} A^0\alpha^T \otimes \beta^T & \text{, falls } z = 0 \wedge z' = 1 \\ A^0\alpha^T \otimes I_m & \text{, falls } z > 0 \wedge z' = z+1 \\ I_n \otimes B^0 & \text{, falls } z = 1 \wedge z' = 0 \\ I_n \otimes B^0\beta^T & \text{, falls } z > 1 \wedge z' = z-1 \\ 0 & \text{sonst.} \end{cases} \tag{3.4}$$

Dabei gilt $A^0 = -Ae$ und $B^0 = -Be$. Ferner sei angemerkt, daß $A^0\alpha^T$ eine $n \times n$–Matrix ist, die aus der Matrixmultiplikation der $n \times 1$–Matrix A^0 und des Vektors bzw. der $1 \times n$–Matrix α^T entsteht. Auf dieselbe Weise entsteht die $m \times m$–Matrix $B^0\beta^T$. I_n bzw. I_m ist die Identitätsmatrix der Dimension n bzw. m. Die ersten beiden Fälle der Gleichung 3.4 repräsentieren die Ankunft eines neuen Kunden und unterscheiden, ob die Station unmittelbar vor der Ankunft leer ist oder nicht. Die Ankunft in der leeren Station hat den Beginn eines neuen Bedienprozesses zur Konsequenz, wohingegen die Ankunft in der nicht–leeren Station den vorgefundenen Bedienprozeß nicht beeinflußt. Die beiden letzten Fälle korrespondieren mit dem Bedienende eines Kunden. Auch dabei ist wiederum zu unterscheiden, ob die Station unmittelbar nach dem Bedienende leer ist oder nicht. Lediglich im Fall der nicht–leeren Station beginnt ein neuer Bedienprozeß. Aufgrund des gewählten Abstraktionsniveaus des Zustandsraumes existieren neben den in Gleichung 3.4 definierten Zustandsübergängen zusätzlich zustandsinterne Übergänge, die jedoch für den Betrachter unsichtbar bleiben. Zustandsinterne Übergänge zeichnen sich durch einen Phasenwechsel des Ankunfts– oder des Bedienprozesses aus, ohne jedoch die Ankunft eines neuen Kunden oder das Bedienende eines Kunden auszulösen. Zustandsinterne Transitionen werden durch die Matrizen $Q[z,z], \forall z \in Z$ erfaßt und sind folgendermaßen definiert:

$$Q[z,z] = \begin{cases} A & \text{, falls } z = 0 \\ A \oplus B & \text{sonst.} \end{cases} \tag{3.5}$$

Wiederum nimmt der Makrozustand 0 eine Sonderposition ein, da in der leeren Station offensichtlich kein Bedienprozeß aktiv ist.

Die derartig definierte unendliche Markovkette läßt sich durch ihre Generatormatrix Q kompakt darstellen. Es gilt:

$$Q = \begin{pmatrix} A & A^0\alpha^T \otimes \beta^T & & & \\ I_n \otimes B^0 & A \oplus B & A^0\alpha^T \otimes I_m & & \\ & I_n \otimes B^0\beta^T & A \oplus B & A^0\alpha^T \otimes I_m & \\ & & I_n \otimes B^0\beta^T & A \oplus B & A^0\alpha^T \otimes I_m \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{3.6}$$

Die i–te Zeile bzw. Spalte von Q repräsentiert die Zustandsübergänge aus dem Makrozustand i heraus bzw. in den Makrozustand i hinein. Die Tridiagonalgestalt der Generatormatrix sowie die regelmäßige Struktur, die sich ab der dritten Zeile bzw. Spalte einstellt, läßt erkennen, daß es sich bei der Markovkette um einen QBD handelt (vgl. [73]). Die Analyse dieses QBDs anhand Matrix–geometrischer Methoden liefert die stationäre Grenzverteilung

$\pi = (\pi_0, \pi_1, \pi_2, \dots)^T$ mit folgender Beziehung:

$$\begin{aligned} \pi_i^T &= \pi_1^T R^{i-1} \quad \forall i \geq 1 \\ \pi_0^T e + \sum_{i=1}^{\infty} \pi_i^T e &= \pi_0^T e + \pi_1^T (I-R)^{-1} e = 1 \end{aligned} \tag{3.7}$$

Die Komponente $\pi_i, i \geq 0$ der stationären Grenzverteilung π ist selbst ein Vektor, der die Wahrscheinlichkeit derjenigen Mikrozustände in der stationären Phase des Systems angibt, die durch den Makrozustand i repräsentiert sind. Dementsprechend ist $\pi_i^T e$ die Wahrscheinlichkeit des Makrozustands i in der stationären Phase des Systems. Die Matrix R ergibt sich gemäß Neuts [73] aus der eindeutigen nicht–negativen Lösung der quadratischen Matrix–Gleichung (3.8) derart, daß der Spektralradius von R kleiner als 1 ist.

$$A^0 \alpha^T \otimes \beta^T + R(A \oplus B) + R^2 (I_n \otimes B^0 \beta^T) = 0 \tag{3.8}$$

Verschiedene Verfahren zur numerischen Berechnung der R–Matrix werden in [61] ausführlich erläutert. Anhand der Matrix R lassen sich die Vektoren π_0 und π_1 (und mit Gleichung (3.7) auch alle übrigen Vektoren) aus dem linearen Gleichungssystem

$$\pi Q = 0$$

ermitteln. Für weitere Details sei an dieser Stelle auf [73] verwiesen.

Die Generatormatrix Q sowie der Vektor π bilden die Basis zur Berechnung der im Fokus stehenden Leistungsmaße der PH/PH/1/∞–FCFS Station. Im folgenden werden die Momente der Verteilungen der Kundenpopulation, der Durchlaufzeit und des Abgangsprozesses berechnet.

3.2.1 Momente der Populationsverteilung

Die k–ten Momente $E[N^k]$ der Populationsverteilung errechnen sich aus der folgenden Beziehung:

$$E[N^k] = \sum_{i=0}^{\infty} i^k \pi_i^T e \tag{3.9}$$

Somit ergibt sich der Erwartungswert $E[N]$ zu;

$$\begin{aligned} E[N] &= \sum_{i=0}^{\infty} i \pi_i^T e \\ &= \pi_1^T \sum_{i=1}^{\infty} i R^{i-1} e \\ &= \pi_1^T (I-R)^{-2} e \end{aligned} \tag{3.10}$$

Für das zweite Moment $E[N^2]$ der Populationsverteilung gilt:

$$\begin{aligned} E[N^2] &= \sum_{i=0}^{\infty} i^2 \pi_i^T e \\ &= \pi_1^T R \sum_{i=2}^{\infty} i(i-1) R^{i-2} e + \pi_i^T \sum_{i=1}^{\infty} i R^{i-1} e \\ &= = 2\pi_1^T R (I-R)^{-3} e + E[N] \end{aligned} \tag{3.11}$$

3.2.2 Momente der Durchlaufzeitverteilung

Die Durchlaufzeit eines Kunden der PH/PH/1/∞–FCFS Station entspricht der Zeitspanne, die der Kunde zwischen seiner Ankunft und dem Zeitpunkt des Verlassens in der Station verweilt. Der Erwartungswert $E[D]$ der Durchlaufzeitverteilung ergibt sich mit Little's Gesetz aus dem Erwartungswert $E[N]$ der Populationsverteilung und der Rate λ der Zwischenankunftszeitverteilung zu

$$E[D] = E[N]/\lambda \tag{3.12}$$

Zur Berechnung höherer Momente werde zunächst ein beliebiger Kunde der Station beobachtet. Unmittelbar nach dessen Ankunft befindet sich das Modell in einem gewissen Makrozustand $z \in Z$. Aufgrund der Reihenfolge erhaltenden Bediendisziplin des Bedieners haben weitere Ankünfte offensichtlich keinen Einfluß auf die Durchlaufzeit des betrachteten Kunden. Unter Vernachlässigung des Ankunftsprozesses entspricht die Durchlaufzeit des Kunden somit der Zeit, die bis zum Erreichen des Makrozustands $z'_0 = 0$ vergeht. Folglich läßt sich die Durchlaufzeit anhand der absorbierenden Markovkette bestimmen, die aus der Generatormatrix Q durch Vernachlässigung des Ankunftsprozesses hervorgeht. Der Zustandsraum Z' dieser Markovkette entspricht dem Zustandsraum Z mit dem absorbierenden Zustand z'_0. Die Interpretation eines Makrozustands $z \in Z'$ ist die, daß unmittelbar nach der Ankunft eines neuen Kunden genau z Kunden im System anwesend sind. Zu beachten ist dabei jedoch, daß sich die durch einen Makrozustand $z \in Z'$ repräsentierten Mikrozustände aufgrund der Vernachlässigung des Ankunftsprozesses von den durch den entsprechenden Makrozustand $z \in Z$ repräsentierten Mikrozuständen unterscheiden. Konkret enthält ein Makrozustand $z \in Z'$ mit $z \neq z'_0$ genau m Mikrozustände, die zu den m Phasen des Bedienprozesses korrespondieren. Die Generatormatrix Q' der absorbierenden Markovkette hat die folgende Gestalt:

$$Q' = \begin{pmatrix} 0 & 0 \\ T'^0 & T' \end{pmatrix} \tag{3.13}$$

Ferner haben die Matrix T' und der Vektor T'^0 die Darstellung:

$$T' = \begin{pmatrix} T'_1 & & & & \\ T'_2 & T'_1 & & & \\ & T'_2 & T'_1 & & \\ & & T'_2 & T'_1 & \\ & & & \ddots & \ddots \end{pmatrix} \tag{3.14}$$

mit

$$\begin{array}{lcl} T_1' & = & B \\ T_2' & = & B^0\beta^T \\ T'^0 & = & (-T_1'e, 0, 0, \dots)^T \end{array} \tag{3.15}$$

Der Startvektor der Markovkette gibt die Zustandsverteilung unmittelbar nach der Ankunft eines Kunden an. Die Wahrscheinlichkeit, daß sich das Modell nach einer Ankunft in dem absorbierenden Zustand z_0' ist 0, da insbesondere der neu eingetroffene Kunde selbst in der Station verweilt. Der Startvektor hat somit die Gestalt $(0, \tau)^T$ mit $\tau = (\tau_1, \tau_2, \dots)^T$. Für die Komponenten $\tau_i, i \geq 1$ gilt:

$$\begin{array}{lcl} \tau_1^T & = & \frac{1}{\lambda}\pi_0^T(A^0\beta^T) \\ \tau_i^T & = & \frac{1}{\lambda}\pi_{i-1}(A^0 \otimes I_B), \quad \forall i \geq 2 \end{array}$$

Die k–ten Momente $E[D^k]$ der Absorbtionszeit bzw. der Durchlaufzeitverteilung lassen sich anhand des Resultats von Neuts [73] folgendermaßen berechnen:

$$E[D^k] = (-1)^k k! \tau T^{-k} e \tag{3.16}$$

Auf die Bestimmung des Erwartungswertes kann verzichtet werden, da er bereits aus Little's Gesetz hervorgeht. Dennoch wird zunächst die Inverse T'^{-1} der Matrix T' dargestellt und die daraus abgeleitete Berechnung des Erwartungswertes interpretiert.

$$T'^{-1} = \begin{pmatrix} T_1'^{-1} & & & \\ UT_1'^{-1} & T_1'^{-1} & & \\ U^2T_1'^{-1} & UT_1'^{-1} & T_1'^{-1} & \\ U^3T_1'^{-1} & U^2T_1'^{-1} & UT_1'^{-1} & T_1'^{-1} \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{3.17}$$

mit

$$U = -T_1'^{-1}T_2 \tag{3.18}$$

Mit dem Resultat 3.16 berechnet sich der Erwartungswert der Durchlaufzeit folgendermaßen:

$$\begin{array}{lcl} E[D] & = & -\tau_1^T T_1'^{-1} e \\ & & -\tau_2^T (UT_1'^{-1}e + T_1'^{-1}e) \\ & & -\tau_3^T (U^2T_1'^{-1}e + UT_1'^{-1}e + T_1'^{-1}e) \\ & & -\dots \\ & = & -\sum_{i=1}^{\infty} \tau_i^T \sum_{j=0}^{i-1} U^j T_1'^{-1} e \end{array} \tag{3.19}$$

Die Durchlaufzeit stimmt also mit der bzgl. der τ_i gewichteten Summe der Durchlaufzeit eines neu eintreffenden Kunden überein, der die Station mit bereits $i-1$ anwesenden Kunden vorfindet. Sei also ein derartiger Kunde betrachtet, der die Station mit bereits $i-1$ anwesenden Kunden antrifft. Dann setzt sich seine Durchlaufzeit aus der restlichen Bedienzeit $\tau_i^T T_1'^{-1} e$

des momentan bedienten Kunden, den Bedienzeiten $\tau_i^T U^j T_1'^{-1} e, j = 1, \ldots i-2$ von $i-2$ weiteren Kunden und seiner eigenen Bedienzeit $\tau_i^T U^{i-1} T_1'^{-1} e$ zusammen. Da die Bedienzeiten der Kunden jedoch unabhängig identisch verteilt sind, sind alle Werte $\tau_i^T U^j T_1'^{-1}$ für alle $j \geq 1$ identisch. Diese Tatsache äußert sich darin, daß alle Potenzen U^j für Werte $j \geq 1$ identisch sind, d.h. $U^j = U, \forall j \geq 1$. Konkret besteht die Matrix U aus gleichen Zeilen, die aus dem Vektor β gebildet werden, und $-UT_1'^{-1}e$ ist ein Vektor, dessen Einträge mit der mittleren Bedienzeit der Station übereinstimmen, d.h. $-UT_1'^{-1}e = E[B]e$. Mit diesen Überlegungen besitzt die Matrix T'^{-2} die folgende Gestalt:

$$T'^{-2} = \begin{pmatrix} T_1'^{-2} & & & \\ W_1 & T_1'^{-2} & & \\ W_2 & W_1 & T_1'^{-2} & \\ W_3 & W_2 & W_1 & T_1'^{-2} \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{3.20}$$

mit

$$W_i = UT_1'^{-2} + (i-1)(UT_1'^{-1})^2 + T_1'^{-1}UT_1'^{-1}, \forall i \geq 1$$

Somit hat das zweite Moment der Durchlaufzeitverteilung die Darstellung:

$$\begin{aligned} E[D^2] &= 2\tau_1^T T_1'^{-2} e \\ &\quad +2\tau_2^T (W_1 + T_1'^{-2}) e \\ &\quad +2\tau_3^T (W_1 + W_2 + T_1'^{-2}) e \\ &\quad + \ldots \\ &= 2\sum_{i=1}^{\infty} \tau_i^T \left(\sum_{j=1}^{i-1} W_j + T_1'^{-2} \right) e \end{aligned} \tag{3.21}$$

Einige elementaren Umformungen liefern den folgenden Ausdruck für das zweite Moment der Durchlaufzeitverteilung:

$$\begin{aligned} E[D^2] &= 2\sum_{i=2}^{\infty} (i-1)\tau_i^T UT_1'^{-2} e + \sum_{i=3}^{\infty} (i-1)(i-2)\tau_i^T (UT_1'^{-1})^2 e \\ &\quad +2\sum_{i=2}^{\infty} (i-1)\tau_i^T T_1'^{-1} UT_1'^{-1} e + 2\sum_{i=1}^{\infty} \tau_i^T T_1'^{-2} e \end{aligned} \tag{3.22}$$

Ebenso wie $-UT_1'^{-1}e = E[B]e$ gilt, ist $2UT_1'^{-1}e$ der Vektor, dessen Komponenten mit dem zweiten Moment der Bedienzeitverteilung übereinstimmt, d.h. $2UT_1'^{-1}e = E[B^2]e$. Unter der zusätzlichen Ausnutzung der Definition der τ_i besitzt das zweite Moment der Durchlaufzeit-

verteilung schließlich die Darstellung:

$$\begin{aligned} E[D^2] &= E[B^2]\frac{1}{\lambda}\pi_1(I-R)^{-2}(A^0\otimes I_m)e \\ &\quad +2E[B]^2\frac{1}{\lambda}\pi_1 R(I-R)^{-3}(A^0\otimes I_m)e \\ &\quad -2E[B]\frac{1}{\lambda}\pi_1(I-R)^{-2}(A^0\otimes I_m)T_1'^{-1}e \\ &\quad +2\frac{1}{\lambda}\pi_0(A^0\beta^T)T_1'^{-2}e \\ &\quad +2\frac{1}{\lambda}\pi_1(I-R)^{-1}(A^0\otimes I_m)T_1'^{-2}e \end{aligned} \tag{3.23}$$

3.2.3 Momente des Abgangsprozesses

Das Interesse dieses Abschnitts liegt in der Charakterisierung der Zeitspanne zwischen aufeinanderfolgenden Kundenabgängen der PH/PH/1/∞–FCFS Station. Da in der Station weder Kunden verschwinden noch neue Kunden generiert werden, stimmt die Rate des Abgangsprozesses mit der Rate λ des Ankunftsprozesses überein. Für den Erwartungswert $E[X]$ der Zwischenabgangszeit gilt daher:

$$E[X] = \frac{1}{\lambda} \tag{3.24}$$

Wie bereits im Fall der Durchlaufzeitverteilung lassen sich die höheren Momente $E[X^k]$ durch die Abbildung des Abgangsverhaltens auf eine absorbierende Markovkette bestimmen. Die Konstruktion dieser Markovkette basiert auf den folgenden Überlegungen. Die Station werde unmittelbar nach dem Abgang eines Kunden betrachtet. Dann ist die Verteilung der Zeitspanne bis zu dem Abgang des nächsten Kunden gesucht. Dazu sind zwei Fälle zu unterscheiden. Im ersten Fall ist die Station unmittelbar nach dem betrachteten Kundenabgang nicht leer. Dann hängt die Zeit bis zum nächsten Kundenabgang aufgrund der Reihenfolge erhaltenden Bediendisziplin ausschließlich von der Bedienzeit des nächsten Kunden ab. Sie ist folglich sowohl von dem Ankunftsprozeß als auch von der Anzahl der anwesenden Kunden unabhängig. Im zweiten Fall ist die Station unmittelbar nach dem betrachteten Kundenabgang leer. Die Zeit bis zum nächsten Kundenabgang entspricht dann der Zeit bis zur nächsten Ankunft eines Kunden zzgl. der Bedienzeit desselben Kunden. Der Zustandsraum Z'' dieser (endlichen) absorbierenden Markovkette zur Berechnung der Momente des Abgangsprozesses der PH/PH/1/∞–FCFS Station enthält somit zwei Makrozustände 0 und 1 sowie den absorbierenden Zustand z_0''. Die Makrozustände 0 bzw. 1 repräsentieren die Situation, daß die Station unmittelbar nach einem Kundenabgang leer bzw. nicht–leer ist und enthalten aufgrund der obigen Schilderungen n bzw. m Mikrozustände.

Die Generatormatrix Q'' dieser Markovkette hat die folgende Gestalt:

$$Q'' = \begin{pmatrix} T'' & T''^0 \\ 0 & 0 \end{pmatrix} \tag{3.25}$$

mit

$$T'' = \begin{pmatrix} A & A^0\beta^T \\ 0 & B \end{pmatrix} \tag{3.26}$$

und

$$T''^0 = \begin{pmatrix} 0 \\ B^0 \end{pmatrix} \tag{3.27}$$

Der Startvektor $(0, \tau_0'', \tau_1'')^T$ gibt die Zustandsverteilung unmittelbar nach einem Kundenabgang an. Somit gilt:

$$\tau_0''^T = \frac{1}{\lambda}\pi_1^T(I_n \otimes B^0) \tag{3.28}$$

$$\tau_1''^T = \frac{1}{\lambda}\pi_2^T(I - R)^{-1}(e_n \otimes B^0\beta^T) \tag{3.29}$$

Die k–ten Momente $E[X^k]$ des Abgangsprozesses resultieren wiederum aus dem Resultat von Neuts und dem Vektor $\tau'' = (\tau_0'', \tau_1'')^T$, d.h.

$$E[X^k] = (-1)^k k!\tau''T''^{-k}e \tag{3.30}$$

Die zur Berechnung des zweiten Moments benötigte Matrix T''^{-2} hat die folgende Darstellung:

$$T''^{-2} = \begin{pmatrix} A^{-2} & Y \\ 0 & B^{-2} \end{pmatrix} \tag{3.31}$$

mit

$$\begin{aligned} X &= -A^{-1}(A^0\beta^T)B^{-1} = e_n\beta^T B^{-1} \\ Y &= A^{-1}X + XB^{-1} \end{aligned}$$

Mit dem Resultat 3.30 hat das zweite Moment des Abgangsprozesses die Darstellung:

$$\begin{aligned} E[X^2] &= 2\left(\tau_0''^T A^{-2}e + \tau_0''^T A^{-1}(e_n\beta^T)B^{-1}e + \tau_0''^T(e_n\beta^T)B^{-2}e + \tau_1''^T B^{-2}e\right) \\ &= 2\tau_0''^T A^{-2}e - 2E[B]\tau_0''^T A^{-1}e + E[B^2]||\tau_0''||_1 + 2\tau_1''^T B^{-2}e \end{aligned}$$

Schließlich besitzt der quadratische Variationskoeffizient c_X^2 des Abgangsprozesses die Darstellung:

$$c_X^2 = \frac{E[X^2]}{E[X]^2} - 1 \tag{3.32}$$

3.3 Analysierbare Modellklasse

Durch die Verwendung von Phasenverteilungen ist es gelungen, die Approximation der Durchlaufzeit gemäß Krämer/Langenbach–Belz und die Approximation gemäß Marshall des originären Dekompositionsverfahrens nach Kühn/Whitt in diesen Spezialfall durch exakte Verfahren zu ersetzen. Von den zu Beginn dieses Abschnittes zusammengefaßten Prinzipien des Dekompositionsverfahrens lassen sich somit die Aspekte 3 und 2 durch die folgenden beiden Prinzipien ersetzen:

1. Approximation der Verteilung aller Ankunfts–, Abgangs– und Bedienprozesse durch Phasenverteilungen.
2. Analyse der isolierten Subnetze/Stationen durch Matrix–geometrische Methoden auf Basis ihrer QBD–Darstellung.

Auf diese Weise läßt sich das Dekompositionsverfahren leicht um Stationstypen erweitern, deren dynamisches Verhalten durch QBDs darstellbar ist. Einige dieser Stationstypen werden im folgenden aufgelistet:

- PH/PH/1/∞–FCFS Stationen.
- PH/PH/1/K–FCFS Stationen, die sich durch beschränkte Warteräume mit K Warteplätzen auszeichnen.
- PH/M/m/∞–FCFS bzw. PH/M/m/K–FCFS Stationen mit negativ–exponentiell verteilten Bedienzeiten und m identischen Bedienern.
- Verschiedene Stationen mit Batch–Ankunfts– und/oder Batch–Bedienprozessen.
- Stationen mit MAP–Ankunfts– bzw. Bedienprozessen.

Zudem lassen sich PH/PH/∞ Stationen (Infinite Server) durch mehrere PH/M/m–FCFS Stationen approximieren.

Im folgenden zweiten Teil dieser Arbeit wird erläutert, wie sich das dynamische Verhalten von Fork/Join–Stationen auf QBDs abbilden lassen. Damit ist auch dieser Stationstyp der Analyse mit dem Dekompositionsverfahren zugänglich.

Teil II

Modellierung und Analyse von Fork/Join-Warteschlangennetzen

Kapitel 4

Fork/Join–Netze

Das übergeordnete Ziel dieser Arbeit liegt in der Bereitstellung einer Methode zur Analyse von erweiterten Fork/Join–Warteschlangennetzen. Das dazu herangezogene Dekompositionsverfahren wurde bereits im ersten Teil beschrieben. Ebenso wurde auf die Erweiterung dieser Methode durch die Verwendung von Phasenverteilungen zur Charakterisierung der Ankunft-, Abgangs- und Bedienprozesse der Stationen eingegangen. Die Idee der Analyse besteht in der Abbildung der isoliert betrachteten Stationen des Netzes auf Quasi–Birth–and–Death (QBD) Prozesse, die sich aus mathematischer Sicht sehr angenehm mittels Matrix–geometrischer Methoden lösen lassen.

Dieser zweite Teil widmet sich zunächst der Präzisierung und Definition des betrachteten Typs von Fork/Join–Stationen. Es wird sich jedoch schnell zeigen, daß dieser Stationstyp keine QBD–Struktur besitzt und somit der übliche Lösungsweg versagt. Stattdessen wird in Kapitel 5 ein eingeschränktes Modell betrachtet, das die ursprüngliche Fork/Join–Station hinsichtlich gewisser Analyseresultate nach oben abschätzt. Aufgrund dieser Tatsache wird im folgenden der Terminus Upper–Bound Modell verwandt. Für das Upper–Bound Modell lassen sich die gewohnten Leistungsmaße wie die Momente der Populationsverteilung, der Durchlaufzeitverteilung und die Momente der Verteilung des Abgangsprozesses bestimmen bzw. approximieren. Einige Experimente in Abschnitt 5.2 belegen die sehr hohe Approximationsgüte der Analyseresultate für das Upper–Bound Modell bezogen auf das ursprüngliche Fork/Join Modell. Aufgrund der Kenntnis der Momente des Abgangsprozesses ist die Schnittstelle zur Umgebung des Modells im Sinne des Dekompositionsverfahrens realisiert. Es zeigt sich, daß die Integration des Upper–Bound Modells in das Dekompositionsverfahren in zyklusfreien Fork/Join–Warteschlangennetzen häufig zu erfreulich guten Ergebnissen insbesondere hinsichtlich der Erwartungswerte der Durchlaufzeitverteilungen der isolierten Stationen als auch des gesamten Netzes führt. In zyklischen Netzen ist das Verfahren nur sehr bedingt einsetzbar, da speziell die fehlende Berücksichtigung von Korrelationen in den Ankunftsprozessen zu deutlichen Fehlern führen kann und im allgemeinen auch führen wird.

Die recht simple Struktur der betrachteten Fork/Join–Station schränkt die praktische Relevanz dieses Modells stark ein. Daher stellt das Kapitel 6 eine Methode zur Erweiterung dieser Modellklasse vor, die auf einem Aggregierungsansatz beruht. Die Idee besteht darin, komple-

xe Fork/Join–Strukturen auf die einfache Station abzubilden, in dem die parallelen Stränge durch geeignete Aggregate ersetzt werden. Als Aggregattyp werden PH/PH/1–∞ Stationen genutzt, deren Bedienzeitverteilung derart eingestellt wird, daß die Momente der Durchlaufzeitverteilung unter dem vorgegebenen Ankunftsprozeß mit denen der zu ersetzenden Netze übereinstimmt. Die Klasse der Netztypen, für die sich diese Aggregierungsmethode eignet, wird schließlich im Abschnitt 6.4 definiert.

Der nachfolgende Abschnitt beschreibt zunächst das primäre Modell des in dieser Arbeit betrachteten Typs von Fork/Join–Stationen.

4.1 Das primäre Modell

Der in diesem Kapitel betrachtete Typ von Fork/Join–Stationen ist in Abbildung 4.1 graphisch dargestellt. Ein Netz dieses Typs besitzt einen Eingangsbereich, eine feste Anzahl M unabhängiger paralleler Bediener mit je einem in der Kapazität unbeschränkten Warteraum sowie einen Ausgangsbereich. Die parallelen Bedieneinrichtungen seien willkürlich von 1 bis M durchnumeriert, und $\mathcal{M}$ bezeichne die Menge dieser Indizes, d.h.

$$\mathcal{M} = \{1, \ldots, M\} \tag{4.1}$$

Kunden betreten das Netz in dem Eingangsbereich und verlassen das Netz über den Ausgangsbereich.

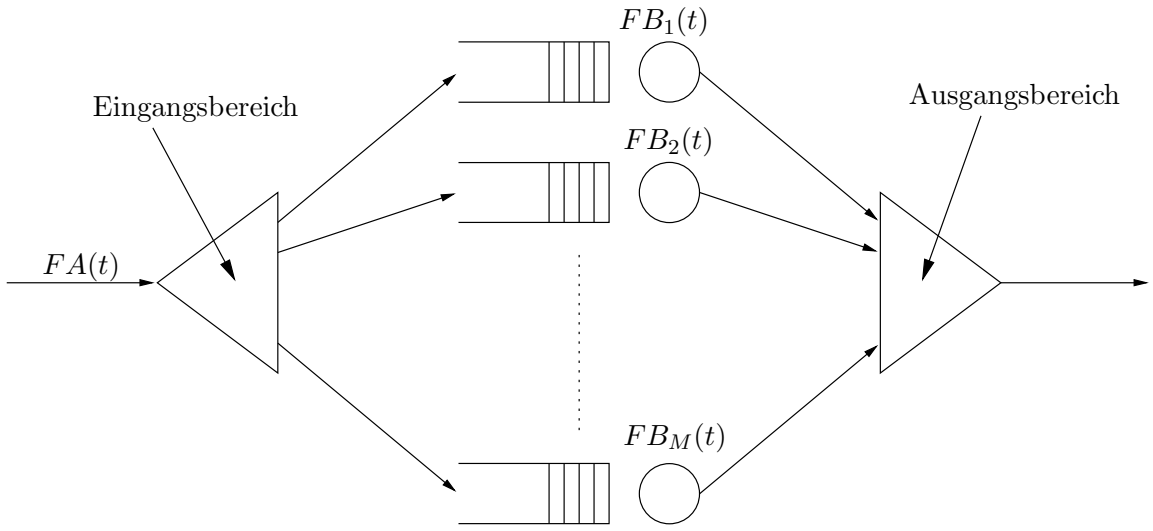

Abbildung 4.1: Graphisches Modell einer Fork/Join–Station

Der zeitliche Abstand, in dem Kunden im Eingangsbereich der Fork/Join–Station eintreffen, sei durch eine kontinuierliche Zufallsvariable mit der Verteilungsfunktion $FA(t)$ charakterisiert. Trifft ein neuer Kunde im Eingangsbereich ein, so richtet dieser Kunde unverzüglich an jeden der M Bediener genau einen Teilauftrag zur Bearbeitung. Anschließend wartet der Kunde im Ausgangsbereich auf die Erfüllung all seiner Teilaufträge. Die Bediener bearbeiten

die an sie gerichteten Aufträge in der Reihenfolge des Eintreffens nacheinander ab (FCFS). Die Bearbeitungszeit von Aufträgen, die an den i–ten Bediener gerichtet werden, ist durch eine kontinuierliche Zufallsvariable mit der Verteilungsfunktion $FB_i(t), \forall i \in \mathcal{M}$ beschrieben. Nachdem alle Teilaufträge eines Kunden vollständig bearbeitet wurden, verläßt der Kunde das Netz ohne weitere Verzögerung.

Den Prinzipien des Dekompositionsverfahrens nach Kühn/Whitt folgend wird auch im Fall der Fork/Join–Station aus Abbildung 4.1 ein Erneuerungs–Ankunftsprozeß vorausgesetzt. Die Ankunftsabstände aufeinanderfolgender Kunden, die durch die Verteilungsfunktion $FA(t)$ beschrieben sind, sind somit unabhängig identisch verteilt. Ferner wird die Verteilungsfunktion $FA(t)$ lediglich bzgl. ihrer Rate und ihres Variationskoeffizienten betrachtet und durch eine spezielle Phasenverteilung (A, α) approximiert (vgl. dazu Anhang A). Ebenso seien die Bedienzeitverteilungen der Aufträge je Bediener unabhängig identisch verteilt. Die Bedienzeitverteilungen werden ebenfalls bzgl. ihrer Rate und ihres quadratischen Variationskoeffizienten betrachtet und durch geeignete Phasenverteilungen mit den Repräsentationen $(B_i, \beta_i), \forall i \in \mathcal{M}$ charakterisiert. Die folgenden Darstellungen gehen davon aus, daß die Phasenverteilungen (A, α) bzw. (B_i, β_i) m_0 bzw. $m_i, \forall i \in \mathcal{M}$ Phasen (transiente Zustände) besitzen.

Das Ziel dieses Kapitels liegt letztendlich in der stationären Analyse der in der Abbildung 4.1 skizzierten Fork/Join–Station. Von speziellem Interesse sind die Verteilungen der Kundenanzahlen, der Durchlaufzeiten der Kunden sowie die Verteilung des Abgangsprozesses. Dem in dieser Arbeit üblichen Vorgehen folgend, wird zu diesem Zweck zunächst ein geeignetes mathematisches Modell der Fork/Join–Station konstruiert. Wie in Kapitel 2 dargelegt, werden insbesondere Markovketten betrachtet, die eine QBD–Struktur aufweisen. Derartige Markovketten lassen sich anhand Matrix–geometrischer Methoden mathematisch besonders angenehm behandeln. Im folgenden wird ein primäres Modell dargestellt, das diese Eigenschaft zunächst nicht besitzen wird. Stattdessen wird es die Basis eines modifizierten Modells bilden, das eine QBD–Gestalt aufweist und geeignet ist, obige Leistungsmaße zu approximieren.

Der Zustandsraum des primären Modells für die Fork/Join–Station aus der Abbildung 4.1 läßt sich auf unterschiedlichen Abstraktionsniveaus definieren, die jeweils unterschiedlich tiefe Einblicke in das Systemverhalten gewähren. Der im folgenden gewählte Zustandsraum Z beschreibt die Fork/Join–Station auf einem recht hohen Abstraktionsniveau:

$$Z := \{z = (n_1, \ldots, n_M) : n_i \geq 0, \forall i \in \mathcal{M}\} \tag{4.2}$$

Bezogen auf das Modell ist die Interpretation eines M–Tupel $(n_1, \ldots, n_M)$ die, daß die n_i die Anzahl der Aufträge an der i–ten Station angibt. Die Wahl dieses Zustandsraums verdeckt die konkreten Phasen, in denen sich der Ankunftsprozeß und die M Bedienprozesse aufgrund ihrer Phasenverteilungen befinden. Die Zustände $z \in Z$ werden aufgrund des gewählten niedrigen Detaillierungsgrades mit dem Begriff *Makrozustände* bezeichnet. Wie bereits in den Ausführungen zur Analyse der PH/PH/1/∞–FCFS Station (vgl. Abschnitt 3.2) erläutert, repräsentiert ein Makrozustand endlich viele *Mikrozustände*, die auf einem höheren Detaillierungsgrad die aktuellen Phasen des Ankunftsprozesses und der Bedienprozesse berücksichtigen.

Da Kunden sich solange in der Fork/Join–Station befinden, bis all ihre M Teilaufträge vollständig bearbeitet sind, gibt die Komponente mit dem maximalen Wert n_i die Gesamtzahl

anwesender Kunden in einem Makrozustand an. Somit ist die Anzahl $N(z)$ der anwesenden Kunden im Makrozustand $z = (n_1, \ldots, n_M)$ folgendermaßen definiert:

$$N(z) := \max_{i \in \mathcal{M}} n_i \tag{4.3}$$

Mit den Repräsentationen (A, α) (m_0 Phasen) bzw. (B_i, β_i) (m_i Phasen) der phasenverteilten Zwischenankunftszeiten bzw. Bedienzeiten sind die Übergänge $Q[z, z']$ zwischen zwei disjunkten Makrozuständen $z = (n_1, \ldots, n_M)$ und $z' = (n'_1, \ldots, n'_M), z, z' \in Z$ wie folgt definiert:

$$Q[z,z'] := \begin{cases} A^0 \alpha^T \otimes \left(\bigotimes_{i=1}^{M} B_i^A(n_i) \right) & , \quad \forall l \in \mathcal{M} : n'_l = n_l + 1 \\ I_{m_0} \otimes \left(\bigotimes_{i=1}^{j-1} K_i(n_i) \right) \otimes & , \quad \exists j \in \{1, \ldots, M\} : n'_j = n_j - 1, \\ \otimes B_j^D(n'_j) \otimes \left(\bigotimes_{i=j+1}^{M} K_i(n_i) \right) & \quad \forall l \in \{1, \ldots, M\} \backslash j : n'_l = n_l \\ 0 & \text{sonst.} \end{cases} \tag{4.4}$$

Der erste Fall der Gleichung 4.4 drückt aus, daß z' aus z aufgrund der Ankunft eines neuen Kunden hervorgeht. An allen Stationen, deren Bediener im Makrozustand z untätig waren, beginnt ein neuer Bedienprozeß. Alle übrigen Bediener setzen den bereits im Makrozustand z gültigen Bedienprozeß fort. Dieses unterschiedliche Systemverhalten in einem Makrozustand z mit mindestens einem untätigen Bediener findet in den Matrizen $B_i^A(z), i = 1, \ldots, M$ folgendermaßen Berücksichtigung:

$$B_i^A(n) := \begin{cases} \beta_i^T & , \text{ falls } n = 0 \\ I_{m_i} & , \text{ falls } n > 0 \end{cases} \tag{4.5}$$

Der zweite Teil der Gleichung 4.4 reflektiert das Bedienende eines Auftrags in der i–ten Station. Warten in dieser Station weitere Aufträge, so beginnt offensichtlich unmittelbar ein neuer Bedienprozeß. Andernfalls ist der Bediener anschließend untätig. Die Matrizen $B_i^D(z), i = 1, \ldots, M$ sind demzufolge definiert:

$$B_i^D(n) := \begin{cases} B_i^0 & , \text{ falls } n = 0 \\ B_i^0 \beta_i^T & \text{sonst.} \end{cases} \tag{4.6}$$

Der Ankunftsprozeß und die übrigen Bedienprozesse bleiben von dem Bedienende an der Station i unberührt. Diese Tatsache drückt sich durch die Identitätsmatrizen I_A und $K_i(z), i = 1, \ldots, M$ in der zweiten Zeile der Gleichung 4.4 aus. Diese Matrizen berücksichtigen ferner, daß evtl. einige der Bediener untätig sind. Die Matrizen $K_i(z)$ sind daher folgendermaßen definiert:

$$K_i(n) = \begin{cases} I_{m_i} & , \text{ falls } n \geq 1 \\ 1 & \text{sonst.} \end{cases} \tag{4.7}$$

Angemerkt sei, daß das Bedienende eines Auftrags nicht zwangsläufig den Abgang eines Kunden zur Folge hat. Dazu muß zusätzlich die Bedingung $N(z') = N(z) - 1$ erfüllt sein.

Aufgrund des gewählten Abstraktionsniveaus des Zustandsraums Z blieben in den bisherigen Betrachtungen solche Systemveränderungen unberücksichtigt, die aus Phasenwechseln im Ankunftsprozeß oder in den Bedienprozessen resultieren, ohne jedoch ein Bedienende oder eine Kundenankunft auszulösen. Diese internen, nach außen nicht sichtbaren Systemveränderungen drücken sich in zustandsinternen Übergängen folgendermaßen aus:

$$Q[z,z] = A \oplus \left(\bigoplus_{i=1}^{M} B_i^I(n_i) \right) \tag{4.8}$$

In einem Makrozustand z ist ein Phasenwechsel in einem der Bedienprozesse nur dann möglich, wenn der entsprechende Bediener auch tatsächlich einen Auftrag bearbeitet. Dies drückt sich in den Matrizen B_i^I aus:

$$B_i^I(n) = \begin{cases} 0 & \text{, falls } n = 0 \\ B_i & \text{, falls } n > 0 \end{cases} \tag{4.9}$$

Dabei benutzt die Gleichung 4.8 die Definition $C \oplus 0 = C$.

Da der Zustandsraum Z offensichtlich in N Dimensionen unbeschränkt ist, besitzt das durch 4.2 und 4.4 definierte Modell keine QBD–Struktur (vgl. [73]). Insbesondere läßt sich keine Partitionierung des Zustandsraums Z in endliche Teilzustände derart finden, daß Zustandsübergänge nur innerhalb einer Partition oder zwischen direkt benachbarten Partitionen möglich sind. Durch eine geschickte Einschränkung des Zustandsraums lassen sich jedoch Modelle beschreiben, die sich auf QBDs abbilden lassen und die das Verhalten der in der Abbildung 4.1 dargestellten Fork/Join–Station sehr gut approximieren. Zwei solche Modelle haben S. Balsamo, L. Donatiello und N. van Dijk in der Arbeit [11] entwickelt. Die Modelle sind derart konzipiert, daß sie die Berechnung einer oberen und einer unteren Schranke für die mittlere Kundenpopulation und die mittlere Durchlaufzeit erlauben. Gemäß [11] werden die Modelle mit den Begriffen *Upper–Bound Modell* und *Lower–Bound Modell* bezeichnet. Die Idee des Upper–Bound Modells liegt darin, die Differenz der momentanen Auftragsanzahlen zwischen je zwei parallelen Stationen durch Schranken $U_{i,j}$ nach oben zu begrenzen. In Zuständen, in denen eine Schranke $U_{i,j}$ erreicht ist, wird der j–te Bediener, dessen Bedienende zur Verletzung der Schranke $U_{i,j}$ führte, blockiert. Durch eine geschickte Partitionierung des Zustandsraums läßt sich die QBD–Struktur dieses Modells erkennen. Hierauf wird im folgenden Abschnitt genauer eingegangen. Das Lower–Bound Modell basiert auf der Idee, die Warteräume von $M-1$ der M parallelen Stationen in ihrer Kapazität zu begrenzen. Teilaufträge, die an einer Station keinen freien Warteplatz vorfinden, werden abgewiesen bzw. gehen verloren. Der Zustandsraum dieses Modells ist in nur einer Dimension unbeschränkt und läßt sich daher auf einen QBD abbilden.

In dieser Arbeit wird das Upper–Bound Modell zur Berechnung oberer Schranken herangezogen und im folgenden Kapitel vorgestellt. Die Begründung für diese Wahl liegt darin, daß in vielen Anwendungsgebieten, insbesondere im Bereich Logistik, der Verlust von Teilaufträgen keine adäquate Modellierung ist.

Kapitel 5

Das Upper–Bound Modell

Dieser Abschnitt stellt ein modifiziertes formales Modell des in Abbildung 4.1 skizzierten Fork/Join–Netzes vor. Dieses Modell besitzt die Eigenschaft, daß es sich auf einen QBD abbilden läßt und somit der Analyse anhand Matrix–geometrischer Methoden zugänglich ist. Die Analyse wird zur Approximation interessierender Leistungsmaße der Fork/Join–Station herangezogen. Im Fokus steht insbesondere die Ermittlung der Verteilungen der Kundenpopulation, der Durchlaufzeit und des Abgangsprozesses des stationären Systems bzgl. ihrer Momente. Hinsichtlich der Erwartungswerte der Kundenpopulation und der Durchlaufzeit wird dieses Modell eine obere Schranken für die entsprechenden Maße des Fork/Join–Netzes aus Abbildung 4.1 liefern. In Anlehnung an die Arbeit [11], in der dieses Modell vorgestellt wurde, wird im folgenden die Bezeichnung *Upper–Bound Modell* genutzt.

Das Upper–Bound Modell entsteht aus dem in Abschnitt 4.1 definierten formalen Modell durch eine geschickte Einschränkung des Zustandsraums Z (vgl. Gl. 4.2). Aus Sicht des realen Modells besteht die Idee darin, die Anzahl der momentan an einer der parallelen Stationen verweilenden Aufträge in Abhängigkeit der Auftragsanzahlen der übrigen Stationen zu begrenzen. Im formalen Modell läßt sich diese Restriktion durch ganzzahlige positive Konstanten $U(n,m)$ realisieren, die einen Makrozustand $z = (n_1, \ldots, n_M) \in Z$ genau dann auch als Makrozustand des Upper–Bound Modells markieren, wenn die Bedingung $n_i - n_j \leq U(n_i, n_j) =: U_{i,j}, \forall i,j = 1, \ldots, M$ erfüllt ist. Der Zustandsraum $\tilde{Z}$ des Upper–Bound Modells ist somit folgendermaßen definiert:

$$\tilde{Z} = \{(n_1, \ldots, n_M) \in Z : -U_{ji} \leq n_i - n_j \leq U_{ij}, 1 \leq i,j \leq M\} \tag{5.1}$$

Diese Einschränkung läßt sich durch die Betrachtung der Auswirkungen auf die beiden möglichen Ereignistypen einer Kundenankunft und des Bedienendes an einem der Bediener noch deutlicher herausstellen. Dazu sei ein beliebiger gültiger Makrozustand $z = (n_1, \ldots, n_M) \in \tilde{Z}$ des Upper–Bound Modells betrachtet. Die Ankunft eines neuen Kunden im Modell bewirkt einen Wechsel in den Makrozustand $z' = (n'_1, \ldots, n'_M)$. z' hat die Eigenschaft, daß all seine Zustandskomponenten einen um genau 1 höheren Wert haben, als die entsprechenden Komponenten des Makrozustands z, d.h. $n'_i = n_i + 1, \forall i \in \{1, \ldots, M\}$. Da für z die Bedingung $n_i - n_j \leq U_{i,j}, \forall i,j = 1, \ldots, M$ erfüllt ist, ist sie es offensichtlich auch für z', und somit ist z' ebenfalls ein gültiger Makrozustand von $\tilde{Z}$. Die Ankunft eines neuen Kunden kann somit

niemals zu einer Verletzung der gewählten Restriktionen führen.

Seien also die Auswirkungen obiger Restriktionen auf das Bedienende eines Bedieners betrachtet. Dazu wird wiederum ein gültiger Makrozustand $z = (n_1, \ldots, n_M) \in \tilde{Z}$ des Upper–Bound Modells herangezogen. z habe zusätzlich die Eigenschaft, daß für mindestens ein Paar $i, j \in \{1, \ldots, M\} : n_i > n_j > 1$ die Beziehung $n_i - n_j = U_{i,j} - 1$ erfüllt ist. Das Bedienende des momentan von der j-ten Station bearbeiteten Auftrags führt zu einer Erhöhung der Differenz $n_i - n_j$ um 1 und somit in den Makrozustand $z' = (n'_1, \ldots, n'_M)$ mit $n'_i - n'_j = U_{i,j}$. Ein weiteres Bedienende an der Station j im Makrozustand z' hätte unmittelbar eine Verletzung der Schranke $U_{i,j}$ zur Folge. Zur Auflösung dieser Problematik sieht die Semantik des Upper–Bound Modells vor, daß im Fall $n'_i - n'_j = U_{i,j}$ der j–te Bediener seine Bedientätigkeit einstellt bzw. blockiert. Somit kann kein Bedienende–Ereignis von dieser Station ausgelöst werden. Der j–te Bediener setzt seine Tätigkeit unmittelbar zu dem Zeitpunkt fort, beginnt also mit der Bedienung des nächsten Auftrags, wenn die kritische Situation durch ein Bedienende an der i–ten Station aufgelöst wird. Etwas salopp formuliert kann ein Bediener sich also nur selbst in eine kritische Situation bringen, wenn er nämlich zu schnell arbeitet, und benötigt dann die Hilfe eines anderen Bedieners, um aus dieser Situation wieder herauszufinden. Aus diesen Ausführungen wird klar, daß eine derartige kritische Situation nur in solchen Makrozuständen des Upper–Bound Modells eintreten kann, in denen für mindestens ein Paar i, j die Beziehung $n_i - n_j = U_{i,j}$ gilt.

Die Zustandsübergänge $\tilde{Q}[z, z']$ für Makrozustände $z, z' \in \tilde{Z}$ des Upper–Bound Modells sind auf der Grundlage der Zustandsübergänge $Q[z, z']$ des primären Modells geeignet anzupassen. Für disjunkte Makrozustände $z = (n_1, \ldots, n_M)$ und $z' = (n'_1, \ldots, n'_M)$ $(z \neq z', z, z' \in \tilde{Z})$ wirken sich obige Erläuterungen in folgenden Modifikationen der Zustandsübergänge $Q[z, z']$ aus:

$$\tilde{Q}[z,z'] := \begin{cases} A^0 \alpha^T \otimes \left(\bigotimes_{i=1}^{M} \tilde{B}_i^A(z) \right) & , \quad \forall l \in \{1,\ldots,M\} : n'_l = n_l + 1 \\ I_{m_0} \otimes \left(\bigotimes_{i=1}^{j-1} \tilde{K}_{i,j}(z) \right) \otimes & \\ \otimes \tilde{B}_j^D(z') \otimes \left(\bigotimes_{i=j+1}^{M} \tilde{K}_{i,j}(z) \right) & , \quad \begin{array}{l} \exists j \in \{1,\ldots,M\} : n'_j = n_j - 1, \\ \forall l \in \{1,\ldots,M\} \backslash j : n'_l = n_l \end{array} \\ 0 & \text{sonst.} \end{cases} \tag{5.2}$$

Der erste Fall der Gleichung 5.2 repräsentiert wiederum einen Zustandswechsel aufgrund der Ankunft eines neuen Kunden. Im primären Modell unterscheiden die Matrizen $B_i^A(n)$ die unterschiedlichen Auswirkungen einer Ankunft auf die leere $(n = 0)$ bzw. nicht–leere $(n > 0)$ Station i. Im Upper–Bound Modell müssen die Matrizen $\tilde{B}_i^A(z)$ den zusätzlichen Fall der blockierten Station i berücksichtigen. Sie sind daher folgendermaßen definiert:

$$\tilde{B}_i^A(z) = \begin{cases} \beta_i^T & \text{, falls } n_i = 0 \wedge \forall j \in \{1,\ldots,M\} \backslash i : n_j - n_i < U_{i,j} \\ I_{m_i} & \text{, falls } n_i > 0 \wedge \forall j \in \{1,\ldots,M\} \backslash i : n_j - n_i < U_{i,j} \\ 1 & \text{sonst.} \end{cases} \tag{5.3}$$

Der zweite Fall der Gleichung 5.2 erfaßt Zustandsübergänge aufgrund des Bedienendes an einem der M parallelen Bediener. Die Schilderungen zu Beginn dieses Abschnitts haben verdeutlicht, daß eine Blockierung der i–ten Station nur durch ein Bedienende an derselben Station i eintreten kann. Andererseits kann die Blockierung nur aufgrund eines Bedienendes an einer anderen Station aufgehoben werden. Folglich müssen die Matrizen $\tilde{B}_i^D(z)$ die möglichen Übergänge aus nicht–blockierenden Zuständen in blockierende Zustände erfassen. Umgekehrt berücksichtigen die Matrizen $\tilde{K}_{i,j}(z)$ die möglichen Übergänge aus blockierenden in nicht–blockierende Makrozustände. Folglich sind die Matrizen $\tilde{B}_i^D(z)$ und $\tilde{K}_{i,j}(z)$ auf der Grundlage der Matrizen $B_i^D(n)$ und $K_i(n)$ des primären Modells (vgl. Gl. 4.6 und 5.5) folgendermaßen zu definieren:

$$\tilde{B}_i^D(z') := \begin{cases} B_i^0 & \text{, falls } n_i' = 0 \vee \exists j \in \{1,\ldots,M\}\backslash i : n_j' - n_i' = U_{i,j} \\ B_i^0 \beta_i^T & \text{, sonst} \end{cases} \tag{5.4}$$

Die erste Zeile dieser Gleichung repräsentiert den Fall, daß das Bedienende eines Auftrags an der Station i in einen Zustand mit der leeren Station i führt oder in einen Zustand, in dem die Station i blockiert ist.

$$\tilde{K}_{i,j}(z) = \begin{cases} \beta_i^T & \text{, falls } n_i > 0 \wedge n_j - n_i = U_{j,i} \wedge \forall l \in \{1,\ldots,M\}\backslash\{i,j\} : n_l - n_i < U_{l,i} \\ I_{m_i} & \text{, falls } n_i \geq 1 \wedge \forall l \in \{1,\ldots,M\}\backslash i : n_l - n_i < U_{l,i} \\ 1 & \text{sonst.} \end{cases} \tag{5.5}$$

Die erste Zeile der Definition der Matrizen $\tilde{K}_{i,j}(z)$ reflektiert den Fall, daß der im Makrozustand z blockierte Bediener der Station i aufgrund eines Bedienendes an der Station j seine Tätigkeit wieder aufnimmt und mit der Bedienung des nächsten Auftrags beginnt.

Die zustandsinternen Übergänge, die aus dem gewählten Abstraktionsniveau und den Phasenrepräsentationen der Zwischenankunftszeiten und der Bedienzeiten resultieren, haben für Zustände $z \in \tilde{Z}$ die folgende Gestalt:

$$\tilde{Q}[z,z] = A \oplus \left(\bigoplus_{i=1}^{M} \tilde{B}_i^I(z) \right) \tag{5.6}$$

Im primären Modell unterscheidet die Matrix $B_i^I(n), i = 1,\ldots,M$ die Fälle, daß die i-te Station leer ist ($n = 0$) oder nicht ($n \geq 0$). Im Upper–Bound Modell wird diese Unterscheidung um den Fall erweitert, daß in einem Makrozustand $z \in \tilde{Z}$ die i-te Station nicht leer ist, der Bediener jedoch aufgrund einer Blockierung untätig ist. Somit gilt für die Matrizen $\tilde{B}_i^I(z)$:

$$\tilde{B}_i^I(z) = \begin{cases} 0 & \text{, falls } n_i = 0 \vee \exists j \in \{1,\ldots,M\}\backslash i : n_j - n_i = U_{j,i} \\ B_i & \text{, sonst.} \end{cases} \tag{5.7}$$

Um die für QBDs notwendige Tridiagonalgestalt der Generatormatrix $\tilde{Q}$ des Upper–Bound Modells zu erkennen bzw. herzustellen, wird der Zustandsraum $\tilde{Z}$ in disjunkte Teilzustandsräume $\tilde{Z}_k, k \geq 0$ partitioniert, die die Eigenschaft

$$\tilde{Z} = \bigcup_{k \geq 0} \tilde{Z}_k \tag{5.8}$$

besitzen. Dabei umfaßt die Partition $\tilde{Z}_k$ genau die Zustände $z \in \tilde{Z}$ mit mindestens k Aufträgen an allen parallelen Bedienern und der Zusatzeigenschaft, daß mindestens eine Station genau k Aufträge aufweist. Formal lassen sich die Partitionen $\tilde{Z}_k$ folgendermaßen definieren:

$$\tilde{Z}_k \quad = \quad \{(n_1, \ldots, n_M) \in \tilde{Z} : \min_{1 \le i \le M} n_i = k\}, \quad k \ge 0 \tag{5.9}$$

Diese Partitionierung hat die für QBDs geforderte Eigenschaft, daß Zustandsübergänge entweder intern, also innerhalb einer Partition $\tilde{Z}_k$, oder zwischen zwei direkt benachbarten Partitionen $\tilde{Z}_{k-1}$ und $\tilde{Z}_k$ auftreten. Die Abbildung 5.1 veranschaulicht diese Partitionierung anhand eines Fork/Join–Netzes mit zwei parallelen Bedienern und den Schranken $U_{1,2} = U_{2,1} = 2$. Die Zwischenankunftszeiten sind negativ–exponentiell mit dem Parameter λ und die Bedienzeiten negativ–exponentiell mit den Raten μ_1 und μ_2 verteilt.

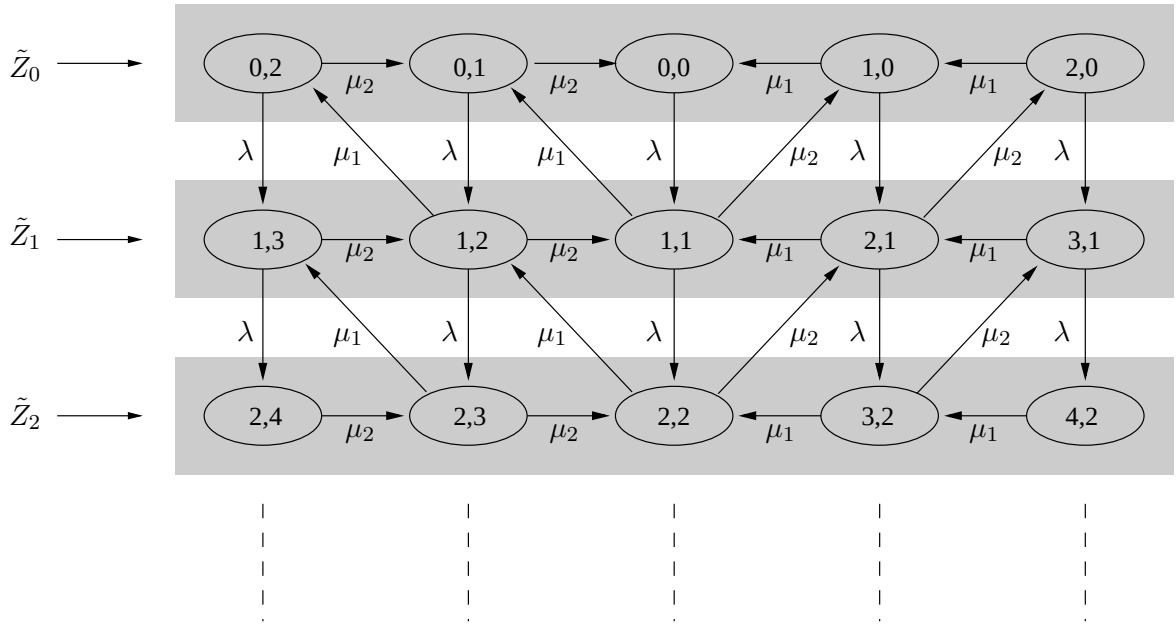

Abbildung 5.1: Zustandsübergänge des Upper Bound Modells

Mit der Partitionierung des Zustandsraums $\tilde{Z}$ in die Partitionen $\tilde{Z}_k$ sowie mit einer fest gewählten Ordnung auf der Menge der Makrozustände innerhalb der Partitionen besitzt die Generatormatrix $\tilde{Q}$ des Upper–Bound Modells die in (5.10) dargestellte Tridiagonalgestalt.

$$\tilde{Q} = \begin{pmatrix} \tilde{T}_1 & \tilde{T}_0 & & & \\ \tilde{T}_2 & \tilde{S}_1 & \tilde{S}_0 & & \\ & \tilde{S}_2 & \tilde{S}_1 & \tilde{S}_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{5.10}$$

Die Matrizen $\tilde{T}_0$, $\tilde{T}_1$ und $\tilde{T}_2$ reflektieren das abnorme Übergangsverhalten in Makrozustände mit keinem oder genau einem Auftrag an einer der Stationen (vgl. Gl. (5.3), (5.4)). Die Matrizen $\tilde{T}_0$ und $\tilde{S}_0$ repräsentieren Zustandsübergänge, die zu einer Kundenankunft korrespondieren. Etwas ungewöhnlich an der obigen Zustandspartitionierung ist die Tatsache, daß

Kundenabgänge stets zu Zustandsübergängen innerhalb einer Partition führen. Die entsprechenden Transitionen finden sich also in den Diagonalmatrizen $\tilde{T}_1$ und $\tilde{S}_1$ wieder und nicht, wie in vielen anderen Fällen, in den rückwärtsgerichteten Matrizen $\tilde{T}_2$ und $\tilde{S}_2$. Die rückwärtsgerichteten Matrizen repräsentieren all die Zustandsübergänge, in denen ein Bedienende an einer Station mit geringster Auftragsanzahl eintritt.

In [11] wird gezeigt, daß dieser QBD genau dann positiv rekurrent ist und somit eine stationäre Zustandsverteilung besitzt, wenn $\lambda < \min_{1 \leq i \leq N} \mu_i$ gilt. Dabei ist λ die Rate des Ankunftsprozesses, und die $\mu_i, i = 1, \ldots, M$ sind die Raten der Bedienzeitverteilungen der parallelen Bediener.

Die stationäre Zustandsverteilung $\pi = (\pi_0, \pi_1, \pi_2, \ldots)^T$ des Upper–Bound Modells bestimmt sich aus der Lösung des linearen Gleichungssystems

$$\pi^T \tilde{Q} = 0 \tag{5.11}$$

mit der Normierungsbedingung $\pi e = 1$. Die Komponenten $\pi_i, i \geq 0$ korrespondieren mit der $(i+1)$–ten Zeile der Matrix $\tilde{Q}$. Das Gleichungssystem 5.11 läßt sich anhand Matrix–geometrischer Methoden lösen (vgl. [73, 61]). Das Ergebnis sind die Vektoren π_0 und π_1 sowie die Matrix R, aus der sich die übrigen Vektoren ergeben:

$$\pi_i^T = \pi_1^T R^{i-1}, \quad \forall i \geq 1 \tag{5.12}$$

5.1 Analyse des Upper–Bound Modells

Der Vektor π bildet die Grundlage zur weiteren Analyse des Upper–Bound Modells. Im folgenden wird zunächst die Berechnung der Verteilung der Kundenpopulation gemäß [11] reflektiert. Ferner ergibt sich mit Little's Gesetz aus dem Erwartungswert der Kundenpopulation unmittelbar der Erwartungswert der Durchlaufzeit. Über die Arbeit [11] hinausgehend wird zudem in Abschnitt 5.36 das zweite Moment der Durchlaufzeitverteilung und in Abschnitt 5.1.3 das zweite Moment der Zwischenabgangszeiten für das Upper–Bound Modell approximiert bzw. exakt berechnet. Das erste Moment der Zwischenabgangszeiten entspricht offensichtlich dem Kehrwert der Ankunftsrate des Upper–Bound Modells.

5.1.1 Momente der Kunden–Populationsverteilung

Zur Bestimmung der Momente der Kunden–Populationsverteilung sei zunächst nochmals an die Definition der Kundenanzahl $N(z)$, die sich in einem Makrozustand $z = (n_1, \ldots, n_M)$ im System befindet, erinnert. Demnach entspricht die Kundenanzahl dem Wert der Zustandskomponente i mit dem maximalen Wert n_i (vgl. Gl. 4.3). Da also offensichtlich innerhalb einer Partition $\tilde{Z}_k, k \geq 0$ die Kundenanzahl nicht für alle Makrozustände konstant ist, enthalte der Vektor γ in der j–ten Komponente die Kundenanzahl des bzgl. der gewählten Ordnung j–ten Makrozustands der Partition $\tilde{Z}_0$. Ferner entstehe der Vektor $\gamma^k, k \geq 0$ aus γ durch komponentenweises Potenzieren mit k. Dann ist $(ie + \gamma)$ der entsprechende Populationsvektor

der i–ten Partition. Damit ergibt sich der Erwartungswert $E[N]$ der Kundenpopulation des Upper–Bound Modells im stationären Fall zu:

$$\begin{aligned} E[N] &= \sum_{i=0}^{\infty} \pi_i^T (ie + \gamma) \\ &= \pi_0^T \gamma + \pi_1^T \sum_{i=1}^{\infty} iR^{i-1} e + \pi_1^T \sum_{i=1}^{\infty} R^{i-1} \gamma \\ &= \pi_0^T \gamma + \pi_1^T (I-R)^{-1} \gamma + \pi_1^T (I-R)^{-2} e \end{aligned} \tag{5.13}$$

Das zweite Moment der Verteilung der Kundenpopulation ist:

$$\begin{aligned} E[N^2] &= \sum_{i=0}^{\infty} \pi_i^T (i^2 e + 2i\gamma + \gamma^2) \\ &= \pi_0^T \gamma^2 + 2\pi_1^T R(I-R)^{-3} e + \pi_1^T (I-R)^{-2} e \\ &\quad + 2\pi_1^T (I-R)^{-2} \gamma + \pi_1^T (I-R)^{-1} \gamma^2 \end{aligned} \tag{5.14}$$

Über die Arbeit von Balsamo et a. [11] hinausgehend ergibt sich das k–te Moment der Kundenpopulation allgemeiner zu:

$$E[N^k] = \pi_0^T \gamma^k + \pi_1^T \sum_{i=1}^{\infty} \left(R^{i-1} \sum_{j=0}^{k} \binom{k}{j} i^{k-j} \gamma^j \right). \tag{5.15}$$

5.1.2 Momente der Durchlaufzeitverteilung

Dieser Abschnitt erläutert die Bestimmung der Momente der Durchlaufzeitverteilung des Upper–Bound Modells. Die Durchlaufzeit entspricht der Zeitspanne zwischen der Ankunft eines Kunden in dem Fork/Join–Netz und dem Zeitpunkt, zu dem derselbe Kunde das System verläßt. Mit Little's Gesetz ergibt sich der Erwartungswert $E[D]$ der Durchlaufzeitverteilung aus dem Erwartungswert $E[N]$ der Populationsverteilung und der Rate λ der Zwischenankunftszeitverteilung zu

$$E[D] = E[N]/\lambda \tag{5.16}$$

Eine ähnliche Beziehung zwischen der Populationsverteilung und der Durchlaufzeitverteilung für höhere Momente ist leider nicht bekannt. Im folgenden wird daher ein alternativer Weg beschritten, der zum Ziel hat, das Verhalten des Upper–Bound Modells unmittelbar nach der Ankunft eines Kunden bis zum Verlassen des Netzes auf eine absorbierende Markovkette abzubilden. Die Verteilung der Absorbtionszeit dieser Markovkette entspricht der Durchlaufzeitverteilung des Upper–Bound Modells. Es wird sich zeigen, daß die exakte Bestimmung der Momente der Absorbtionszeit nicht möglich ist. Die Begründung dafür liegt darin, daß im Falle des Upper–Bound Modells die Wahrscheinlichkeiten, mit denen innerhalb des QBD von einem Block der Generatormatrix in den vorherigen gesprungen wird, nicht identisch verteilt sind. Dies erschwert die Analyse der Absorbtionszeit der Markovkette, die zur Ermittlung der Durchlaufzeitverteilung herangezogen wird. Um das zweite Moment der Durchlaufzeit

dennoch zu approximieren, wird die Annahme getroffen, daß im Falle mehrerer Rückwärtsschritte obige Bedingung wiederum approximativ erfüllt ist. Im folgenden wird zunächst die Konstruktion der absorbierenden Markovkette beschrieben und anschließend ihre Analyse erläutert.

Zur Konstruktion der absorbierenden Markovkette werde zunächst ein beliebiger Kunde des Fork/Join–Netzes betrachtet. Dieser Kunde findet das System unmittelbar nach seiner Ankunft in einem gewissen Makrozustand $z \in \tilde{Z}$ vor. Aufgrund der FCFS–Bediendisziplin der M parallelen Bediener haben spätere Ankünfte offensichtlich keinen Einfluß auf die Durchlaufzeit des betrachteten Kunden. Unter Vernachlässigung des Ankunftsprozesses entspricht die Durchlaufzeit somit genau dem Zeitintervall bis zum Erreichen des vollständig leeren System, d.h. bis zum Erreichen des Zustands $z'_0 = (0, \ldots, 0)$.

Das so beschriebene Modellverhalten läßt sich direkt auf eine unendliche absorbierende Markovkette abbilden. Der Zustandsraum $\tilde{Z}'$ dieser Markovkette entspricht dem Zustandsraum $\tilde{Z}$ des QBDs für das Upper–Bound Modell, d.h.

$$\tilde{Z}' = \tilde{Z} \tag{5.17}$$

Der Zustand $z'_0 = (0, \ldots, 0)$ ist absorbierend. Zu beachten ist hierbei, daß sich aufgrund der Vernachlässigung des Ankunftsprozesses die durch einen Makrozustand $z \in \tilde{Z}' = \tilde{Z}$ repräsentierten Mikrozustände in der absorbierenden Markovkette von denen des QBDs für das Upper–Bound Modell unterscheiden. Die Zustandsübergänge $\tilde{Q}'[z, z']$ für Zustände $z, z' \in \tilde{Z}'$ entstehen aus den Zustandsübergängen des QBDs für das Upper–Bound Modell durch Vernachlässigung des Ankunftsprozesses. Für disjunkte Makrozustände $z = (n_1, \ldots, n_M) \in \tilde{Z}'$ und $z' = (n'_1, \ldots, n'_M) \in \tilde{Z}'$ folgt somit:

$$\tilde{Q}'[z, z'] = \begin{cases} \left(\bigotimes_{i=1}^{j-1} \tilde{K}_{i,j}(z)\right) \otimes \tilde{B}^D_j(z') \otimes & \\ \otimes \left(\bigotimes_{i=j+1}^{M} \tilde{K}_{i,j}(z)\right) & , \quad \exists j \in \{1, \ldots, M\} : n'_j = n_j - 1, \\ & \quad \forall l \in \{1, \ldots, M\} \backslash j : n'_l = n_l \\ 0 & \text{sonst.} \end{cases} \tag{5.18}$$

Da in dem Upper–Bound Modell ausschließlich Ankunftsereignisse aus dem Zustand $(0, \ldots, 0)$ herausführen, ist damit bereits die absorbierende Eigenschaft von z'_0 gewährleistet. Die Vernachlässigung des Ankunftsprozesses führt in den zustandsinternen Übergängen $\tilde{Q}'[z, z]$ für Zustände $z \in \tilde{Z}'$ zu folgender Modifikation:

$$\tilde{Q}'[z, z] = \begin{cases} \bigoplus_{i=1}^{M} \tilde{B}^I_i(z) & \text{, falls } z \neq z'_0 \\ 0 & \text{sonst.} \end{cases} \tag{5.19}$$

Die Matrizen $\tilde{K}_{i,j}(z)$, $\tilde{B}^D_j(z)$ und $\tilde{B}^I_i(z)$ sind entsprechend der Matrizen des Upper–Bound Modells definiert (Gl. (5.4), (5.5), (5.7)).

Zur Erkennung der Diagonalgestalt der Generatormatrix $\tilde{Q}'$ der absorbierenden Markovkette wird der Zustandsraum $\tilde{Z}'$ wiederum geeignet partitioniert. Die Partitionierung entspricht der

Partitionierung des Upper–Bound Modells (vgl. Gl. 5.9) unter Ausschluß des absorbierenden Zustands z'_0. Somit gilt:

$$\tilde{Z}'_0 = \tilde{Z}_0 \backslash z'_0 \tag{5.20}$$

$$\tilde{Z}'_k = \tilde{Z}_k, \forall k \geq 1 \tag{5.21}$$

$$\tilde{Z}' = \bigcup_{k=0}^{\infty} \tilde{Z}'_k \cup \{z'_0\} \tag{5.22}$$

Mit der gewählten Partitionierung besitzt die Generatormatrix $\tilde{Q}'$ der absorbierenden Markovkette die folgende Gestalt:

$$\tilde{Q}' = \begin{pmatrix} 0 & \mathbf{0} \\ \tilde{T}'^0 & \tilde{T}' \end{pmatrix} \tag{5.23}$$

Die erste Zeile bzw. Spalte der Matrix repräsentiert den absorbierenden Zustand z'_0. Die Matrix $\tilde{T}'$ hat folgende Diagonalgestalt:

$$\tilde{T}' = \begin{pmatrix} \tilde{T}'_1 & & & & \\ \tilde{T}'_2 & \tilde{S}'_1 & & & \\ & \tilde{S}'_2 & \tilde{S}'_1 & & \\ & & \tilde{S}'_2 & \tilde{S}'_1 & \\ & & & \ddots & \ddots \end{pmatrix} \tag{5.24}$$

Die i–te Zeile bzw. Spalte von $\tilde{T}'$ korrespondiert zu der $(i-1)$–ten Partition $\tilde{Z}'_{i-1}$. Der Vektor $\tilde{T}'^0$ repräsentiert alle Zustandsübergänge, die in den absorbierenden Zustand z'_0 führen. Solche $\tilde{Q}'[z, z'_0]$ zeichnen sich dadurch aus, daß z in genau einer Komponente den Wert 1 besitzt und in allen übrigen Komponenten ausschließlich 0–Einträge hat. Somit erfolgen Transitionen in den absorbierenden Zustand ausschließlich aus Makrozuständen der Partition $\tilde{Z}'_0$ heraus. $\tilde{T}'^0$ hat damit die Gestalt:

$$\tilde{T}'^0 = \begin{pmatrix} -\tilde{T}'_1 e \\ 0 \\ 0 \\ \vdots \end{pmatrix} \tag{5.25}$$

Der Startvektor $(0, \tilde{\tau}')^T$ der absorbierenden Markovkette gibt die Zustandsverteilung unmittelbar nach einer Kundenankunft an. Da sich insbesondere der neu eingetroffenen Kunde selbst im System befindet, jeder der M parallelen Bediener somit mindestens einen Teilauftrag besitzt, sind die Wahrscheinlichkeiten für die Zustände der Partition $\tilde{Z}'_0$ (und für den absorbierenden Zustand z'_0) 0. Der Vektor $\tilde{\tau}'$ hat somit die Darstellung $\tilde{\tau}' = (0, \tilde{\tau}'_1, \tilde{\tau}'_2 \ldots .)$. Die $\tilde{\tau}'_i$ ergeben sich aus der stationären Grenzverteilung π und gewissen Matrizen $\tilde{T}'_0$ und $\tilde{S}'_0$ zu:

$$\tilde{\tau}'_1 = \frac{1}{\lambda} \pi_0 \tilde{T}'_0 \tag{5.26}$$

$$\tilde{\tau}'_i = \frac{1}{\lambda} \pi_{i-1} \tilde{S}'_0 \tag{5.27}$$

Die Matrizen $\tilde{T}'_0$ und $\tilde{S}'_0$ entstehen aus den Matrizen $\tilde{T}_0$ und $\tilde{S}_0$ (vgl. Gl. (5.10)) durch die Projektion aller (Mikro-) Zustandsübergänge aufgrund einer Ankunft, die die Bedienprozesse

unverändert lassen und lediglich einen Phasenwechsel des Ankunftsprozesse zurfolge haben, auf einen einzigen (Mikro-) Zustandsübergang. Die Projektion wirkt sich in einer geeigneten Modifikation der Matrizen der (Makro-) Zustandsübergänge aus. Konkret enthält $\tilde{T}_0'$ Zustandsübergänge aus den Makrozuständen der Partition $\tilde{Z}_0'$ in Makrozustände der Partition $\tilde{Z}_1'$ folgendermaßen:

$$\tilde{T}_0'[z,z'] = \begin{cases} A^0 \otimes \left(\bigotimes_{i=1}^{M} \tilde{B}_i^A(z) \right) & ,\forall l \in \{1,\ldots,M\} : n_l' = n_l + 1 \\ 0 & \text{sonst.} \end{cases} \tag{5.28}$$

Die durch $\tilde{S}_0'$ repräsentierten Zustandsübergänge aus Makrozuständen der Partitionen $\tilde{Z}_k'$ in Makrozustände der Partitionen $\tilde{Z}_{k+1}', \forall k \geq 1$ sind ebenso definiert. Der Unterschied zu der Darstellung von $\tilde{T}_0'$ liegt in der verschiedenen Definition der Matrizen $\tilde{B}_i^A(z)$ für Makrozustände aus $z \in \tilde{Z}_0'$ und $z \in \tilde{Z}_k', k \geq 1$ (vgl. Gl. (5.3)).

Anhand der Matrix $\tilde{Q}'$ und des Vektors $(0, \tilde{\tau}')^T$ ist die absorbierende Markovkette zur Berechnung der Durchlaufzeitverteilung des Upper–Bound Modells vollständig beschrieben. Die Abbildung 5.2 verdeutlicht die Konstruktion anschaulich anhand eines einfachen Beispiels mit zwei parallelen Bedienern und den Schranken $U_{1,2} = U_{2,1} = 2$. Der Ankunftsprozeß und die beiden Bedienprozesse sind negativ–exponentiell mit den Raten λ und μ_1 bzw. μ_2 verteilt.

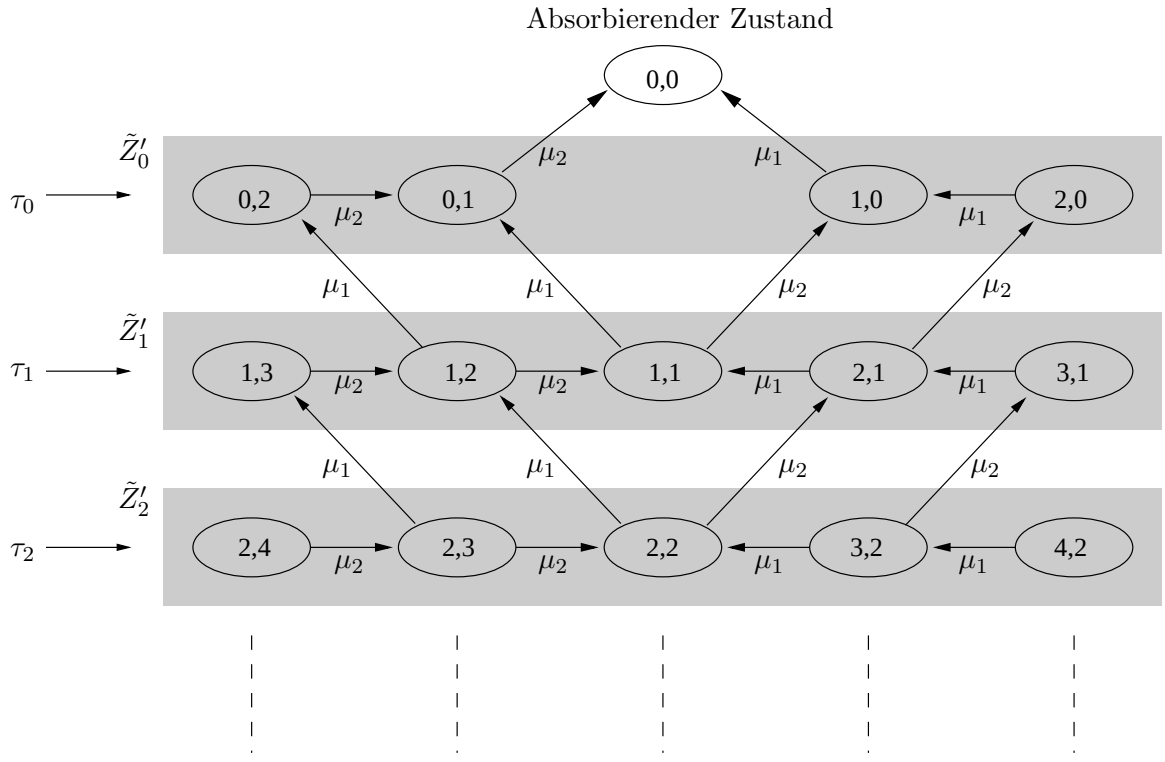

Abbildung 5.2: Absorbierende Markovkette für die Durchlaufzeit

Zur Lösung der durch $\tilde{T}'$ und $\tilde{\tau}'$ beschriebenen absorbierenden Markovkette wird das bereits mehrfach in dieser Arbeit verwendete Resultat von Neuts [73] herangezogen. Demnach

resultieren die k–ten Momente der Absorbtionszeit bzw. der Durchlaufzeitverteilung V der Kunden des Upper–Bound Modells aus folgender Identität:

$$E[D^k] = (-1)^k k! \tilde{\tau}'^T \tilde{T}'^{-k} e \tag{5.29}$$

Die in der Gleichung 5.29 benötigte Inverse $\tilde{T}'^{-1}$ der Matrix $\tilde{T}'$ hat die in Gleichung 5.30 dargestellte Form.

$$\tilde{T}'^{-1} = \begin{pmatrix} \tilde{T}_1'^{-1} & & & & \\ W & \tilde{S}_1'^{-1} & & & \\ UW & U\tilde{S}_1'^{-1} & \tilde{S}_1'^{-1} & & \\ U^2W & U^2\tilde{S}_1'^{-1} & U\tilde{S}_1'^{-1} & \tilde{S}_1'^{-1} & \\ U^3W & U^3\tilde{S}_1'^{-1} & U^2\tilde{S}_1'^{-1} & U\tilde{S}_1'^{-1} & \tilde{S}_1'^{-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{5.30}$$

Die Matrizen W und U sind folgendermaßen definiert:

$$W = -\tilde{S}_1'^{-1}\tilde{T}_2'\tilde{T}_1'^{-1} \tag{5.31}$$

$$U = -\tilde{S}_1'^{-1}\tilde{S}_2' \tag{5.32}$$

Zur Bestimmung des zweiten Moments der Durchlaufzeitverteilung wird im folgenden zunächst, wie auch schon im Falle der in Abschnitt 3.2 behandelten PH/PH/1/∞–FCFS Station, die Darstellung des Erwartungswertes angegeben. Mit der Gleichung 5.29 ergibt sich für den Erwartungswert $E[D]$ der Durchlaufzeitverteilung die folgende Darstellung:

$$\begin{aligned} E[D] &= -\tilde{\tau}_1'^T(We + \tilde{S}_1'^{-1}e) \\ &\quad -\tilde{\tau}_2'^T(UWe + (U+I)\tilde{S}_1'^{-1}e) \\ &\quad -\tilde{\tau}_3'^T(U^2We + (U^2+U+I)\tilde{S}_1'^{-1}e) \\ &\quad -\tilde{\tau}_4'^T(U^3We + (U^3+U^2+U+I)\tilde{S}_1'^{-1}e) \\ &\quad -\cdots \\ &= -\sum_{i=1}^{\infty} \tilde{\tau}_i'^T \left(U^{i-1}We + \sum_{j=0}^{i-1} U^j \tilde{S}_1'^{-1}e \right) \end{aligned} \tag{5.33}$$

Die Vektoren $\tilde{\tau}_i', i \geq 1$ geben die Wahrscheinlichkeiten dafür an, daß ein neu eintreffender Kunde das Modell unmittelbar nach seiner Ankunft in einem der Makrozustände der i–ten Partition $\tilde{Z}_i'$ vorfindet. Seine Durchlaufzeit setzt sich dann aus der restlichen Verweilzeit in der i–ten Partition, der Summe der Verweilzeiten in den Partitionen $\tilde{Z}_{i-1}', \ldots, \tilde{Z}_1'$ und der Verweilzeit in der Partition $\tilde{Z}_0'$ zusammen. Die restliche Verweilzeit in der Partition $\tilde{Z}_i'$ beträgt $\tilde{\tau}_i'\tilde{S}_1'^{-1}e$, die Verweilzeiten in den Partitionen $\tilde{Z}_j', j = i-1, \ldots, 1$ sind $\tilde{\tau}_i'U^j\tilde{S}_1'^{-1}e$, und schließlich beträgt die Verweilzeit in der Partition $\tilde{Z}_0'$ $\tilde{\tau}_i'U^{i-1}We$. Aus diesen Ausführungen geht eine gewisse Analogie zu den Resultaten des Kapitels 3.2.2 über die Analyse der PH/PH/1/∞–FCFS Station hervor. Gedanklich läßt sich das Upper–Bound Modell somit als Single–Server Station mit reihenfolgeerhaltender Bediendisziplin interpretieren. Diesem Gedankenspiel folgend, entsprechen die Zustände des Single–Server Modells den Partitionen des Upper–Bound Modells. Die Besonderheit in der Berechnung der Durchlaufzeitverteilung der PH/PH/1/∞–FCFS Station liegt in der speziellen Struktur des zugrundeliegenden QBDs. Unabhängig vom

Anfangszustand beginnt jeder Block (mit Ausnahme des ersten) mit identischen Verteilungen. Diese Tatsache äußert sich darin, daß die Potenzen U^i für alle $i \geq 1$ identisch sind (vgl. Gl. (3.18) und die nachfolgenden Ausführungen). Der hier betrachtete QBD zur Berechnung der Durchlaufzeitverteilung des Upper–Bound Modells weist keine derartige spezielle Struktur auf, so daß die Potenzen U^i aus Gl. (5.32) nicht identisch sind. Allerdings ist davon auszugehen, daß bei Ankunft eines neuen Kunden im Makrozustand k der Einfluß der Startverteilung in dem entsprechenden Block auf die Startverteilungen in den nachfolgend eingenommenen Blöcken kontinuierlich abnimmt. Nach einer bestimmten Anzahl an Schritten ist dieser Einfluß praktisch vernachlässigbar. Das bedeutet jedoch, daß die Potenzen U^i für große Werte von i konvergieren und die folgende Approximation

$$U^{l+i} = U^l, \forall i \geq 0 \tag{5.34}$$

für eine gewisse Schranke l gerechtfertigt ist. Damit läßt sich dann das zweite Moment $E[D^2]$ der Durchlaufzeitverteilung des Upper–Bound Modells approximieren. Dazu wird zunächst die Matrix T^{-2} benötigt. T^{-2} hat die folgende Gestalt:

$$\tilde{T}^{-2} = \begin{pmatrix} \tilde{T}_1'^{-2} & & & & \\ W_1 & \tilde{S}_1'^{-2} & & & \\ W_2 & U_1 & \tilde{S}_1'^{-2} & & \\ W_3 & U_2 & U_1 & \tilde{S}_1'^{-2} & \\ W_4 & U_3 & U_2 & U_1 & \tilde{S}_1'^{-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{5.35}$$

mit

$$\begin{aligned} W_i &= U^{i-1} W \tilde{T}_1'^{-1} + \sum_{j=0}^{i-1} U^j \tilde{S}_1'^{-1} U^{i-j-1} W, \forall i \geq 1 \\ U_i &= \sum_{j=0}^{i} U^j \tilde{S}_1'^{-1} U^{i-j} \tilde{S}_1'^{-1}, \forall i \geq 1. \end{aligned}$$

Das zweite Moment $E[D^2]$ der Durchlaufzeitverteilung ergibt sich somit aus Gleichung 5.29 zu:

$$\begin{aligned} E[D^2] &= 2\tilde{\tau}_1^T (W_1 e + \tilde{S}_1'^{-2} e) \\ &+ 2\tilde{\tau}_2^T (W_2 e + U_1 e + \tilde{S}_1'^{-2} e) \\ &+ 2\tilde{\tau}_3^T (W_3 e + U_2 e + U_1 e + \tilde{S}_1'^{-2} e) \\ &+ \cdots \\ &= \underbrace{2 \sum_{i=1}^{\infty} \tilde{\tau}_i^T W_i e}_{A} + \underbrace{2 \sum_{j=1}^{\infty} \sum_{i=j+1}^{\infty} \tilde{\tau}_i^T U_j e}_{B} + \underbrace{2 \sum_{i=1}^{\infty} \tilde{\tau}_i^T \tilde{S}_1'^{-2} e}_{C}. \end{aligned} \tag{5.36}$$

Dabei sind A, B und C durch folgende Gleichungen gegeben:

$$
\begin{aligned}
A &= \frac{2}{\lambda}\pi_0^T T_0 \left(W T_1'^{-1} + S_1'^{-1} W\right) e \\
&\quad + \frac{2}{\lambda}\pi_K^T (I-R)^{-1} S_0 \left(U^K W T_1'^{-1} e + S_1'^{-1} U^K W e + U^K S_1'^{-1} W e\right) \\
&\quad + \frac{2}{\lambda}\pi_{2K}^T (I-R)^{-2} S_0 U^K S_1'^{-1} U^K W e \\
&\quad + \frac{2}{\lambda}\sum_{i=1}^{K-1} \pi_i^T S_0 \left(U^i W T_1'^{-1} e + S_1'^{-1} U^i W e + U^i S_1'^{-1} W e\right) \\
&\quad + \frac{2}{\lambda}\sum_{i=1}^{K-1} \pi_{K+i}^T (I-R)^{-1} S_0 \left(U^K S_1'^{-1} U^i W e + U^i S_1'^{-1} U^K W e\right) \\
&\quad + \frac{2}{\lambda}\sum_{j=1}^{K-1}\sum_{i=1}^{K-1} \pi_{i+j}^T S_0 U^j S_1'^{-1} U^i W e
\end{aligned} \tag{5.37}
$$

$$
\begin{aligned}
B &= \frac{2}{\lambda}\pi_K^T (I-R)^{-2} S_0 \left(S_1'^{-1} U^K S_1'^{-1} e + U^K S_1'^{-2} e\right) \\
&\quad + \frac{2}{\lambda}\pi_{2K}^T (I-R)^{-3} S_0 U^K S_1'^{-1} U^K S_1'^{-1} e \\
&\quad + \frac{2}{\lambda}\sum_{i=1}^{K-1} \pi_i^T (I-R)^{-1} S_0 \left(S_1'^{-1} U^i S_1'^{-1} e + U^i S_1'^{-2} e\right) \\
&\quad + \frac{2}{\lambda}\sum_{i=1}^{K-1} \pi_{K+i}^T (I-R)^{-2} S_0 \left(U^K S_1'^{-1} U^i S_1'^{-1} e + U^i S_1'^{-1} U^K S_1'^{-1} e\right) \\
&\quad + \frac{2}{\lambda}\sum_{j=1}^{K-1}\sum_{i=1}^{K-1} \pi_{i+j}^T (I-R)^{-1} S_0 U^j S_1'^{-1} U^i S_1'^{-1} e
\end{aligned} \tag{5.38}
$$

$$
\begin{aligned}
C &= \frac{2}{\lambda}\pi_0^T T_0 S_1'^{-2} e \\
&\quad + \frac{2}{\lambda}\pi_1^T (I-R)^{-1} S_0 S_1'^{-2} e.
\end{aligned} \tag{5.39}
$$

In der Veröffentlichung [11] zeigen Balsamo et al., daß die mittlere Durchlaufzeit $E[D]$ und die mittlere Population $E[N]$ des Upper–Bound Modells obere Schranken für die entsprechenden Werte des primären, in der Abbildung 4.1 skizzierten, Fork/Join–Netzes bilden. Auf die detaillierte Reflexion dieses Beweises wird in dieser Arbeit verzichtet. Informell läßt sich die Aussage jedoch sehr leicht begründen. Dazu werde ein beliebiger Kunde des Fork/Join–Netzes betrachtet. Die Behandlung dieses Kunden durch das primäre Fork/Join–Netz stimmt im wesentlichen mit dem durch das Upper–Bound Modell abgebildete Verhalten überein. Lediglich in dem Fall, daß aufgrund einer oder mehrerer der Restriktionen $U_{i,j}$ einer der Bediener blockiert, d.h. trotz wartender Aufträge untätig bleibt, unterscheidet sich das Modellverhalten von dem realen System. Der entsprechende Bediener des realen Systems unterbricht seine

Tätigkeit nicht. Folglich ist die Durchlaufzeit des Upper–Bound Modell mindestens so groß wie die des primären Netzes und damit eine obere Schranke. Aufgrund des Resultats von Little ist daher auch die mittlere Population des Upper–Bound Modells eine obere Schranke für die mittlere Population des ursprünglichen Fork/Join–Netzes.

5.1.3 Der Abgangsprozeß des Upper–Bound Modells

Dieser Abschnitt erläutert die Bestimmung des Abgangsprozesses des Upper–Bound Modells. Dem Abgangsprozeß kommt in dieser Arbeit eine besondere Bedeutung zu, da er aus technischer Sicht die Integration eines Fork/Join–Netzes in das Dekompositionsverfahren nach Kühn/Whitt zur Analyse allgemeiner Warteschlangennetze erlaubt. Dazu sei nochmals an die Beschreibung des Dekompositionsverfahrens aus Kapitel 2 erinnert. Dort wird die Zerlegung eines Warteschlangennetzes in disjunkte Subnetze als zentrales Prinzip herausgestellt. Die Ankunfts– und Abgangsprozesse bilden die Schnittstelle, über die die Subnetze mit der Umgebung interagieren. Ein weiteres Prinzip ist die Charakterisierung der Schnittstellenprozesse ausschließlich über ihre Erwartungswerte und quadratischen Variationskoeffizienten.

Das Ziel dieses Abschnitts besteht folglich in der Analyse des Upper–Bound Modells bzgl. der ersten beiden Momente des Abgangsprozesses. Konkreter ausgedrückt liegt das Interesse in der Charakterisierung der Zeitspanne aufeinanderfolgender Kundenabgänge. Wie bereits im Fall der Durchlaufzeitverteilung ist zur Bestimmung des Erwartungswertes des Abgangsprozesses keine Arbeit zu leisten. Da in dem Upper–Bound weder Kunden verschwinden noch neue Kunden erzeugt werden, entspricht die Rate des Abgangsprozesses der Rate λ des Ankunftsprozesses. Somit gilt für den Erwartungswert $E[X]$ der Zwischenabgangszeiten die Beziehung:

$$E[X] = \frac{1}{\lambda} \tag{5.40}$$

Zur Bestimmung höherer Momente bietet sich wiederum die Abbildung des Abgangsverhaltens auf eine (endliche) absorbierende Markovkette an. Dazu sei das System unmittelbar nach einem Kundenabgang betrachtet. Gesucht ist die Verteilung der Zeitdauer bis zum nächsten Abgang. Befinde sich das System also zum Beobachtungszeitpunkt in einem Makrozustand $z \in \tilde{Z}$. Dann sind die Fälle $z = (0, \ldots, 0)$ (vollständig leere Fork/Join–Station) und $z \neq (0, \ldots, 0)$ (mindestens ein Kunde in der Fork/Join–Station) zu unterscheiden. Im Falle der leeren Station muß vor dem nächsten Kundenabgang offensichtlich zunächst eine weitere Ankunft stattfinden. Anschließend ist das System nicht leer und kann wie die übrigen Fälle behandelt werden. Sei also angenommen, es befinde sich mindestens ein Kunde in der Fork/Join–Station. Dann hängt die Zeit bis zum nächsten Abgangsereignis ausschließlich von den (restlichen) Bedienzeiten der aktuellen Aufträge an all den Stationen mit einer maximalen Auftragsanzahl n_{max} ab. Der konkrete Wert von n_{max} als auch die Bedienprozesse der übrigen Stationen bleiben hinsichtlich des nächsten Kundenabgangs ohne Einfluß. Insbesondere können somit auch weitere Kundenankünfte vernachlässigt werden. Der Zustandsraum $\tilde{Z}''$ der endlichen Markovkette mit dem absorbierenden Zustand z_0'' läßt sich somit folgendermaßen notieren:

$$\tilde{Z}'' = \{(n_1, \ldots, n_M) : n_i \in \{0, 1\}, \forall i \in \mathcal{M}\} \cup \{z_0''\} \tag{5.41}$$

Der Zustand $(0, \ldots, 0)$ repräsentiert die leere Fork/Join–Station. Die Interpretation der übrigen Zustände ist die, daß all die Stationen, deren Zustandskomponenten den Wert 1 auf-

weisen, dieselbe (maximale) Auftragsanzahl besitzen und alle sonstigen Stationen mit den Zustandskomponenten 0 geringere Auftragsanzahlen besitzen. Übergänge in den absorbierenden Zustand z_0'' sind aus Zuständen mit genau einem 1–Eintrag möglich. Übergänge in den Zustand $(0,\ldots,0)$ sind nicht möglich, so daß dieser Zustand ausschließlich in der Startphase eingenommen werden kann. Die Abbildung 5.3 verdeutlicht diese Konstruktion anhand eines Beispiels mit zwei parallelen Bedienern, negativ–exponentiell verteilten Bedienzeiten und ebenfalls negativ–exponentiell verteilten Zwischenankunftszeiten. Der obere Teil der Grafik veranschaulicht den ursprünglichen QBD des Upper–Bound Modells, bei dem lediglich die Transitionen aufgrund eines Kundenabgangs in den absorbierenden Zustand z_0'' führen. Im unteren Teil der Grafik ist die beschriebene Vereinfachung anhand des Zustandsraumes $\tilde{Z}''$ dargestellt. Die linke bzw. rechte Spalte des oberen Teils wird auf den Zustand $(0,1)$ bzw. $(1,0)$ des unteren Teils abgebildet. Der Zustand $(0,\ldots,0) \in \tilde{Z}$ entspricht dem Zustand $(0,\ldots,0) \in \tilde{Z}''$, und die übrigen Zustände $(n,\ldots,n) \in \tilde{Z}$ werden auf den Zustand $(1,\ldots,1) \in \tilde{Z}''$ projiziert.

Die Zustandsübergänge der Markovkette für disjunkte Zustände $z = (n_1,\ldots,n_M) \in \tilde{Z}''\backslash\{z_0''\}$ und $z' = (n_1',\ldots,n_M') \in \tilde{Z}''\backslash\{z_0''\}$ sind damit folgendermaßen definiert:

$$\tilde{Q}''[z,z'] = \begin{cases} A^0 \otimes \left(\bigotimes_{i\in\mathcal{M}} \beta_i^T \right) & z = (0,\ldots,0) \wedge z' = (1,\ldots,1) \\ \left(\bigotimes_{i<j:n_i=1} I_{m_i} \right) \otimes B_j^0 \otimes \left(\bigotimes_{i>j:n_i=1} I_{m_i} \right) & \begin{array}{l} \exists j \in \mathcal{M} : n_j' = n_j - 1 \wedge \\ \forall l \in \mathcal{M}\backslash\{j\} : n_l' = n_l \wedge \\ \sum\limits_{i\in\mathcal{M}} n_i > 1 \end{array} \\ 0 & \text{sonst.} \end{cases} \tag{5.42}$$

Die internen Transitionen haben die Darstellung:

$$\tilde{Q}''[z,z] = \begin{cases} A & z = (0,\ldots,0) \\ 0 & z = z_0'' \\ \bigoplus\limits_{i\in M:n_i=1} B_i & \text{sonst.} \end{cases} \tag{5.43}$$

Schließlich ergeben sich für Zustände $z = (n_1,\ldots,n_M) \in \tilde{Z}''\backslash\{z_0''\}$ die Transitionen in den absorbierenden Zustand zu:

$$\tilde{Q}''[z,z_0''] = \begin{cases} B_j^0 & n_i = 1 = \sum\limits_{i\in\mathcal{M}} n_i \\ 0 & \text{sonst.} \end{cases} \tag{5.44}$$

Damit besitzt die Generatormatrix $\tilde{Q}''$ der endlichen absorbierenden Markovkette zur Berechnung des Abgangsprozesses des Upper–Bound Modells die Form:

$$\tilde{Q}'' = \begin{pmatrix} 0 & 0 \\ \tilde{T}''^0 & \tilde{T}'' \end{pmatrix} \tag{5.45}$$

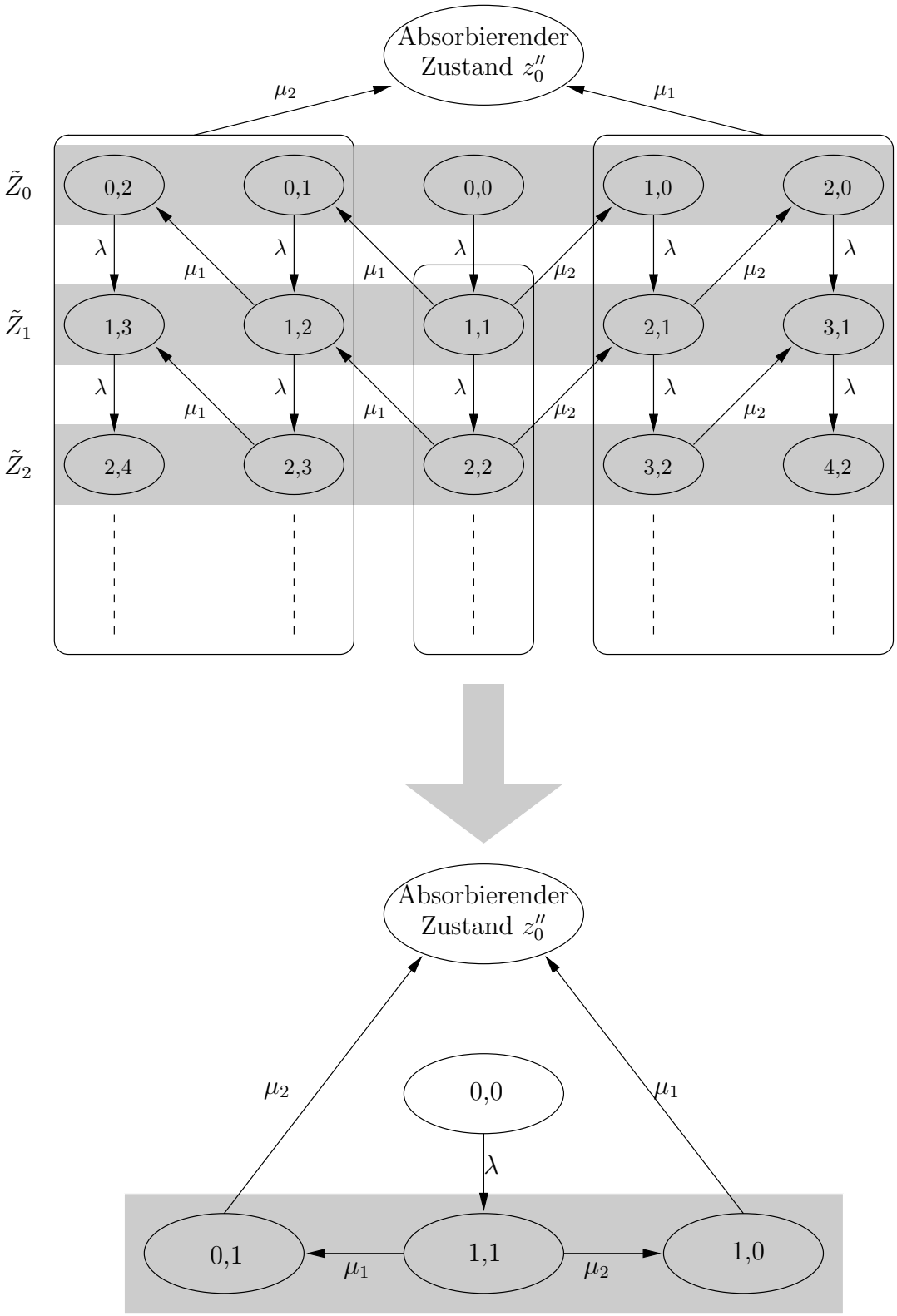

Abbildung 5.3: Absorbierende Markovkette des Abgangsprozesses

Die erste Zeile bzw. Spalte von $\tilde{Q}''$ korrespondiert mit dem absorbierenden Zustand z_0''. Dementsprechend enthält der Vektor $\tilde{T}''^0$ alle Zustandsübergänge in z_0''. Damit haben die Matrix $\tilde{T}''$ und der Vektor $\tilde{T}''^0$ die Gestalt:

$$\tilde{T}'' = \begin{pmatrix} \tilde{T}_1'' & \tilde{T}_0'' \\ 0 & \tilde{S}_1'' \end{pmatrix} \tag{5.46}$$

und

$$\tilde{T}''^0 = \begin{pmatrix} 0 \\ -\tilde{S}_1'' e \end{pmatrix} \tag{5.47}$$

Die erste Zeile bzw. Spalte der Matrix $\tilde{T}''$ und die erste Komponente des Vektors $\tilde{T}''^0$ korrespondieren zu dem Zustand $(0,\ldots,0)$.

Der Startvektor der absorbierenden Markovkette gibt die Verteilung der Zustände $z \in \tilde{Z}''$ unmittelbar nach einem Kundenabgang an. Da die Wahrscheinlichkeit für z_0'' initial 0 ist, hat der Startvektor die Darstellung $(0, \tilde{\tau}'')^T$ mit $\tilde{\tau}'' = (\tilde{\tau}_0'', \tilde{\tau}_1'')^T$. $\tilde{\tau}_0''$ ist die Startwahrscheinlichkeit für den Zustand $(0,\ldots,0)$ und $\tilde{\tau}_1''$ die Startverteilung für die übrigen Zustände. In den Ausführungen zu der Zerlegung des Zustandsraumes $\tilde{Z}$ in die Partitionen $\tilde{Z}_k, k \geq 0$ wurde erläutert, daß Transitionen aufgrund eines Kundenabgangs stets partitionsinterne Übergänge sind. Damit besitzen die Komponenten $\tilde{\tau}_0''$ und $\tilde{\tau}_1''$ die Gestalt:

$$\tilde{\tau}_0'' = r\pi_0\tilde{D}_0 \tag{5.48}$$

$$\tilde{\tau}_1'' = r\left(\pi_0\tilde{D}_1 + \sum_{i=1}^{\infty}\pi_i\tilde{D}_2\right) \tag{5.49}$$

Der Skalar stellt die Normierungsbedingung $\tilde{\tau}_0''^T e + \tilde{\tau}_1''^T e = 1$ sicher. Die Matrizen $\tilde{D}_0, \tilde{D}_1$ und $\tilde{D}_2$ resultieren aus den Zustandsübergängen $\tilde{Q}[z,z'], z, z' \in \tilde{Z}$ (vgl. Gl. (5.2)) aufgrund eines Kundenabgangs durch die Projektion von z' in den Zustandsraum $\tilde{Z}''$. Dabei unterscheiden $\tilde{D}_0$ und $\tilde{D}_1$, ob ein Abgangsereignis innerhalb der Partition $\tilde{Z}_0$ in den Zustand $(0,\ldots,0$ führt oder nicht.

Für Makrozustände $z = (n_1,\ldots,n_M) \in \tilde{Z}_0$ sind die Übergänge der Matrix $\tilde{D}_0$ in den Zustand $z' = (0,\ldots,0) \in \tilde{Z}''$ folgendermaßen definiert:

$$\tilde{D}_0[z,z'] = \begin{cases} I_{m_0} \otimes B_j^0 & \exists j \in \mathcal{M} : n_j = 1 = \sum\limits_{i\in\mathcal{M}} n_i \\ 0 & \text{sonst.} \end{cases} \tag{5.50}$$

Die Zustandsübergänge aus Makrozuständen $z = (n_1,\ldots,n_M) \in \tilde{Z}_0$ aufgrund eines Abgangs, die in Zustände $z' = (n_1',\ldots,n_M') \in \tilde{Z}''\backslash\{(0,\ldots,0), z_0''\}$ führen, sind in der Matrix $\tilde{D}_1$ zusammengefaßt.

$$\tilde{D}_1[z,z'] = \begin{cases} \left(\bigotimes\limits_{\substack{i\in\mathcal{M}:i<j,n_i>0 \\ \forall k\in\mathcal{M}: \\ n_k-n_i<U_{k,i}}} I_{m_i}\right) \otimes B_j^0\beta_j^T \otimes \left(\bigotimes\limits_{\substack{i\in\mathcal{M}:i>j,n_i>0 \\ \forall k\in\mathcal{M}: \\ n_k-n_i<U_{k,i}}} I_{m_i}\right) & \exists j \in \mathcal{M} : n_j > 1 \forall l \in \mathcal{M}\backslash\{j\} : n_l < n_j \\ & \wedge \forall l \in \mathcal{M} : n_l' = 1 \Leftrightarrow n_l \geq n_j - 1 \\ 0 & \text{sonst.} \end{cases} \tag{5.51}$$

Die restlichen Transitionen aufgrund eines Abgangs in der Partition $\tilde{Z}_1$ (die aufgrund obiger Schilderungen auch alle übrigen Partitionen repräsentiert) aus Makrozuständen $z = (n_1, \ldots, n_M) \in \tilde{Z}_1$ in die Zustände $z' \in \tilde{Z}''\backslash\{(0,\ldots,0), z_0''\}$ haben die Darstellung:

$$\tilde{D}_2[z,z'] = \left\{ \left(\bigotimes_{\substack{i\in\mathcal{M}:i<j \\ \forall k\in\mathcal{M}: \\ n_k-n_i<U_{k,i}}} I_{m_i} \right) \otimes B_j^0 \beta_j^T \otimes \left(\bigotimes_{\substack{i\in\mathcal{M}:i>j \\ \forall k\in\mathcal{M}: \\ n_k-n_i<U_{k,i}}} I_{m_i} \right) \right. \quad \begin{array}{l} \exists j \in \mathcal{M} \forall l \in \mathcal{M}\backslash\{j\} : n_l < n_j \wedge \\ \forall l \in \mathcal{M} : n_l' = 1 \Leftrightarrow n_l \geq n_j - 1 \end{array} \tag{5.52}$$

Damit ist die absorbierende Markovkette zur Darstellung des Abgangsverhaltens des Upper–Bound Modells vollständig spezifiziert. Das zweite Moment $E[X^2]$ der Zwischenabgangszeiten ergibt sich mit dem Resultat von Neuts [73] zu:

$$E[X^2] = 2\tilde{\tau}''^T \tilde{T}''^{-2} e \tag{5.53}$$

$\tilde{T}''^{-2}$ hat die Gestalt:

$$\tilde{T}''^{-2} = \begin{pmatrix} \tilde{T}_1''^{-2} & W \\ 0 & \tilde{S}_1''^{-2} \end{pmatrix} \tag{5.54}$$

mit der Matrix

$$W = -\tilde{T}_1''^{-2} \tilde{T}_0'' \tilde{S}_1''^{-1} - \tilde{T}_1''^{-1} \tilde{T}_0'' \tilde{S}_1''^{-2} \tag{5.55}$$

Für $E[X^2]$ ergibt sich damit:

$$E[X^2] = 2\tilde{\tau}_0'' \tilde{T}_1''^{-2} e + 2\tilde{\tau}_0'' W e + 2\tilde{\tau}_1'' \tilde{S}_1''^{-2} e$$

Der Variationskoeffizient c_X des Abgangsprozesses ergibt sich schließlich zu:

$$c_X^2 = \frac{E[X^2]}{E[X]^2} - 1. \tag{5.56}$$

5.1.4 Bestimmung der Schranken U

Mit der Berechnung der Momente der Populationsverteilung, der Durchlaufzeitverteilung und des Abgangsprozesses ist das Analyseverfahren für das Upper–Bound Modell aus technischer Sicht im Sinne dieser Arbeit vollständig beschrieben. Hinsichtlich der Modellbildung bleibt jedoch bisher die Frage offen, wie die Schranken $U_{i,j}$, die das Upper–Bound Modell ausmachen und von dem primären Modell unterscheiden, zu wählen sind. Auf diese Frage werden im folgenden zwei Antworten diskutiert, die sich zum einen aus der Interpretation des Realsystems ergeben und zum anderen aus der Modellanalyse resultieren.

Die erste Variante besteht darin, die Schranken aus der Sicht bzw. der Interpretation des realen Systems heraus zu bestimmen. Diese Wahl bietet sich immer dann an, wenn das Verhalten des realen Systems mit dem des Upper–Bound Modells exakt übereinstimmt bzw. diesem sehr nahe kommt. Das Upper–Bound Modell ist gegenüber dem zu Beginn dieses Kapitels beschriebenen primären Modell in gewisser Weise homogener, da die Schranken $U_{i,j}$ das Vorauseilen eines schnellen Bedieners verhindern. Viele von Menschen geschaffene Systeme besitzen genau diese Eigenschaft. Häufig wird z.B. durch überlagerte Steuerungsprozesse

dafür Sorge getragen, daß parallele Abläufe möglichst gleichmäßig voranschreiten und somit kontrollierbar sind.

Zur Verdeutlichung dieser Aussage sei eine Lieferkette betrachtet, die das Zusammenspiel eines Herstellers und eines Logistikdienstleisters beschreibt. Der unkontrollierte, nicht gesteuerte Lieferprozeß des Logistikdienstleisters, wie er durch das primäre Modell dargestellt würde, führte zu einem erheblichen Lageraufkommen der Zulieferteile. In der Praxis werden jedoch Steuerungsmechanismen eingesetzt, die die bedarfsgerechte Lieferung mit geringen Lagerbeständen oder gar gänzlich ohne Lagerhaltung regeln. Geläufige Fachbegriffe in diesem Kontext sind Just–in–time Lieferung oder spezieller Just–in–sequence Lieferung. Für diesen Fall, daß sich die Schranken $U_{i,j}$ aus dem Verhalten des realen Systems ableiten lassen bzw. daß das Upper–Bound Modell das reale System exakt abbildet, sind offensichtlich auch die errechneten Analyseresultate wie die Momente der Populationsverteilung, der Durchlaufzeitverteilung und des Abgangsprozesses exakt. In diesem Sinne ist diese Variante der Wahl der Werte $U_{i,j}$ sehr angenehm.

Die zweite Möglichkeit bezieht sich auf derartige Fälle, in denen sich die $U_{i,j}$ nicht aus den Eigenschaften des realen Systems heraus ergeben. Wenn das reale System durch das in Abbildung 4.1 skizzierte primäre Modell beschrieben ist, sollte der Approximationsfehler, der aus den Restriktionen des Upper–Bound Modells resultiert, gering sein. In solchen Situationen hilft die Betrachtung des Modells weiter. Dazu sei zunächst daran erinnert, daß die Schranken $U_{i,j}$ die Differenz der Auftragsanzahlen der i–ten und der j–ten Station begrenzen, d.h. für alle Zustände $z = (n_1, \ldots, n_M) \in \tilde{Z}$ gilt: $n_i - n_j \leq U_{i,j}$. Im Fall $n_i - n_j = U_{i,j}, n_j > 0$ ist der j–te Bediener blockiert. Das Ziel muß also darin bestehen, die $U_{i,j}$ so groß zu wählen, daß die stationäre Grenzwahrscheinlichkeit von Zuständen mit blockierten Bedienern gering ist. Diese Aussage läßt sich folgendermaßen formalisieren:

Wähle für ein $\epsilon > 0$ und alle Paare (i, j) mit $i \neq j, i, j \in \{1, \ldots, M\}$ die ganzzahligen Werte $U_{i,j}$ derart, daß für die Zustände $z = (n_1, \ldots, n_M) \in \tilde{Z}$ gilt:

$$\sum_{n_i - n_j = U_{i,j}, n_j > 0} \pi(z) \leq \epsilon \tag{5.57}$$

Die in der Gleichung 5.57 akkumulierten Wahrscheinlichkeiten sind ein Maß dafür, daß das primäre Modell Zustände einnimmt, die in dem Upper–Bound Modell nicht berücksichtigt sind.

Aus technischer Sicht ist die Berechnung der Schranken $U_{1,2}$ und $U_{2,1}$ eines Modells mit zwei parallelen Bedieneinrichtungen leicht realisierbar. Obige Summen entsprechen in diesem Fall der Wahrscheinlichkeit, daß der j-te Bediener blockiert ist. Die Berechnung der Schranken $U_{i,j}$ eines Modells mit $M > 2$ Stationen wird durch die exklusive Betrachtung je zweier Stationen und die Rückführung auf den Fall mit zwei Stationen erreicht.

Die Bestimmung der Schranken $U_{i,j}$ auf diese Art und Weise ist gegenüber der ersten Variante, nämlich der Interpretation des realen System, die zweifelsohne unangenehmere, da sie gewisse Approximationsfehler in den Analyseresultaten in sich birgt. Hinsichtlich dieser Approximationsfehler ist das Analyseverfahren des Upper–Bound Modells zu beurteilen. Ein weiteres Kriterium ist jedoch die Rechenzeit, die das Verfahren benötigt. Die Rechenzeit hängt von

verschiedenen Einflußfaktoren ab, wie z.B. der Größe der $U_{i,j}$, den Phasendarstellungen des Ankunftsprozesses und der Bedienprozesses sowie den Auslastungen der parallelen Bediener.

5.2 Bewertung des Analyseverfahrens

Aus rein technischer Sicht ist mit der Definition und Analyse des Upper–Bound Modells in den Abschnitten 5 und 5.1 ein erster Schritt hinsichtlich der Integration von Fork/Join–Netzes in das Dekompositionsverfahren nach Kühn/Whitt gelungen. Es verbleiben jedoch drei unbeantwortete Aspekte. Diese sind zum einen die Bewertung des Analyseverfahrens, zum anderen die Auswirkungen des Upper–Bound Modells im Kontext eines Warteschlangennetzes und zuletzt die praktische Relevanz des Modells.

Die Auswirkungen, die die Integration des Upper–Bound Modells in ein Warteschlangennetz mit sich bringt, werden in Abschnitt 5.3 beleuchtet. Anhand einiger Experimente wird insbesondere untersucht, ob sich die Eigenschaft des Modells, obere Schranken für die Durchlaufzeit zu liefern, auf die Durchlaufzeit des umgebenden Warteschlangennetzes übertragen läßt.

Auf die praktische Relevanz wird detailliert in Kapitel 6 eingegangen. Dort liegt das zentrale Thema in der Erweiterung des Upper–Bound Modells auf deutlich komplexere Fork/Join–Strukturen, die über die bisherige Betrachtung der Synchronisation von Single–Server Stationen hinausreichen.

Zur Vorbereitung dieser Aspekte beschäftigt sich dieser Abschnitt zunächst mit der Bewertung des vorgestellten Analyseverfahrens. Im Fokus stehen dazu die Kriterien Approximationsgüte der Durchlaufzeit und die Laufzeit des Verfahrens. Eng mit beiden Aspekten verbunden ist die Wahl der Schranken U, die somit ebenfalls ins Zentrum der nachfolgenden Betrachtungen rückt. Das Ziel dieses Abschnittes liegt schließlich darin, eine Vorstellung dieser drei Größen zu vermitteln und damit einhergehend die Einsatzmöglichkeiten bzw. Grenzen des Verfahrens aufzuzeigen.

In den folgenden Ausführungen wird eine Bewertung des Analyseverfahrens anhand verschiedener Experimentreihen vorgenommen. Die Experimentreihen werden insbesondere die in [11] herausgestellte hohe Approximationsgüte des Verfahrens im Falle negativ–exponentiell verteilter Ankunfts– und Bedienprozesse auch im Fall von 1 verschiedener Variationskoeffizienten belegen. Ferner wird experimentell eine gewisse lineare Abhängigkeit der Schranken U von den Variationkoeffizienten herausgearbeitet.

Die Untersuchungen der ersten Serie betrachten ein homogenes Upper–Bound Modell mit zwei Bedienstationen in dem Sinne, daß die Bedienzeiten an beiden Bedienern identische Phasenverteilungen besitzen. Der Ankunftsprozeß bleibt stets unverändert negativ–exponentiell verteilt mit der Rate $\lambda = 1$. Die Experimente unterscheiden sich in der Rate und dem quadratischen Variationskoeffizienten der identischen Bedienprozesse. Im Fokus steht dabei zunächst die Abhängigkeit der Schranken $U_{1,2}$ und $U_{2,1}$ einerseits von der Auslastung der Bediener und andererseits von der Wahl des quadratischen Variationskoeffizienten der Bedienzeitverteilungen. Zudem interessieren der Approximationsfehler bzgl. der Durchlaufzeit und die Rechenzeit des Analyseverfahrens.

5.2.1 Experimentreihe 1: Homogenes Modell mit zwei parallelen Bedienern

In dieser ersten Experimentreihe wird ein homogenes Fork/Join–Netz mit zwei parallelen Bedieneinheiten betrachtet. Dieses Modell ist homogen in dem Sinne, daß die Bedienzeiten der Aufträge an beiden Bedienern identische Verteilungen besitzen. Konkret sind die Bedienzeiten zunächst negativ–exponentiell verteilt. Ebenso sind die Zwischenankunftszeiten des Modells negativ–exponentiell mit der Rate $\lambda = 1$ verteilt. Schließlich ist zur Bestimmung der Schranken $U_{1,2}$ und $U_{2,1}$ des Upper–Bound Modells der Wert $\epsilon = 0.005$ gewählt (vgl. Gleichung 5.57).

In diesem Spezialfall, nämlich im Fall eines homogenen Fork/Join–Netzes mit zwei parallelen Bedienern und negativ–exponentiell verteilten Zwischenankunfts– und Bedienzeiten, läßt sich die mittlere Durchlaufzeit (des primären Modells) exakt angeben [36, 71]. Die Resultate der ersten Experimente sind in der Tabelle 5.1 zusammengefaßt. Dazu wurden die Raten $\mu_1 = \mu_2$ der Bedienzeiten derart gewählt, daß sich für die Bediener Auslastungen ρ zwischen 0.1 und 0.9 ergeben. Die Werte U enthalten die (identischen) Schranken, die sich für das Upper–Bound Modell bei der Wahl $\epsilon = 0.005$ einstellten. Die Spalten 3 und 4 geben die mittlere Durchlaufzeit an, die aus der Analyse des Upper–Bound Modells resultieren (UB) bzw. aus der exakten Analyse des primären Modells (exakt) resultieren. Zudem ist der relative Fehler Δ, um den die Werte des Upper–Bound Modells von den exakten Werten abweichen, angegeben. Zur Berechnung des zweiten Moments bzw. des Variationskoeffizienten der Durchlaufzeit ist kein exaktes Verfahren bekannt. Daher werden die Resultate des Upper–Bound Modells mit der Simulation des primären Modells verglichen. Die Simulation wurde derart eingestellt, daß sich für den Erwartungswert der Durchlaufzeit ein 98%−Konfidenzintervall der Breite 2% einstellte. Schließlich gibt die letzte Spalte die CPU–Zeit an, die das Analyseverfahren des Upper–Bound Modells benötigte. Sämtliche Experimente wurden auf einem AMD Opteron Prozessor mit 2 GHz, 1 MB Cache und 6 GB Hauptspeicher ausgeführt.

		$E[D]$			c_D^2			
ρ	U	UB	Exakt	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
0.1	2	0.1655	0.165278	0.13	0.5673	0.5647	0.46	0
0.2	2	0.3714	0.368750	0.72	0.5850	0.5754	1.67	0
0.3	3	0.6291	0.626786	0.37	0.5913	0.5868	0.77	0
0.4	4	0.9696	0.966667	0.30	0.6017	0.5990	0.45	1
0.5	5	1.4431	1.437500	0.39	0.6145	0.6102	0.70	1
0.6	7	2.1443	2.137500	0.32	0.6258	0.6226	0.51	3
0.7	9	3.3135	3.295833	0.54	0.6415	0.6366	0.77	5
0.8	13	5.6481	5.600000	0.86	0.6585	0.6503	1.26	13
0.9	23	12.7208	12.48750	1.87	0.6794	0.6682	1.68	55

Tabelle 5.1: Homogenes Fork/Join Modell aus zwei Stationen

Die Resultate belegen, daß der Approximationsfehler dieser Experimente sowohl hinsichtlich des Erwartungswertes als auch hinsichtlich des Variationskoeffizienten der Durchlaufzeitverteilung sehr moderat ausfällt. Aufgrund der gegenüber den übrigen Werten größeren Ungenauigkeiten im Falle der Auslastungen $\rho = 0.8$ und insbesondere $\rho = 0.9$ ist zu vermuten, daß

der Approximationsfehler mit steigender Auslastung und konstantem ϵ anwächst.

Ein deutlicherer Trend läßt sich dagegen bzgl. der Schranken U beobachten. Die Werte von U wachsen zunächst etwa linear mit der Auslastung an und weichen von diesem Verhalten erst für Auslastungen $\rho > (\geq)0.7$ ab. Dem Anwachsen von U wird in späteren Experimenten weiterhin nachgegangen. Da die Größe der gewählten Schranken maßgeblich die Größe des QBDs mitbestimmt, der dem Modell zugrundeliegt, steigt die Laufzeit des Analyseverfahrens mit höherer Auslastung deutlich an.

In weiteren Experimenten wurde untersucht, wie sich die Schranken U im Fall eines homogenen Fork/Join–Netzes mit zwei parallelen Bedienern und nicht negativ–exponentiell verteilten Bedienzeiten entwickeln. Die Tabelle 5.2 stellt die Resultate für den Variationskoeffizienten $c_B^2 = 0.5$ der Bedienzeitverteilung dar. Die Bedienraten $\mu_1 = \mu_2$ wurden wiederum derart eingestellt, daß sich für die Bediener unter dem negativ–exponentiell verteilten Ankunftsprozeß mit der Rate $\lambda = 1$ Auslastungen zwischen 0.1 und 0.9 einstellen. In diesem Fall sind weder zur Ermittlung des Erwartungswertes noch zur Ermittlung des zweiten Moments der Durchlaufzeitverteilung des primären Modells exakte Verfahren bekannt, so daß die entsprechenden Werte des Upper–Bound Modells mit den Resultaten einer Simulation verglichen werden.

		$E[D]$			c_D^2			
ρ	U	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
0.1	2	0.1482	0.1482	0.00	0.3145	0.3140	0.16	0
0.2	2	0.3238	0.3232	0.19	0.3451	0.3419	0.94	1
0.3	3	0.5365	0.5365	0.00	0.3744	0.3731	0.35	2
0.4	3	0.8091	0.8073	0.22	0.4123	0.4080	1.05	1
0.5	4	1.1729	1.1713	0.14	0.4525	0.4486	0.87	3
0.6	5	1.7012	1.6990	0.13	0.4990	0.4955	0.71	6
0.7	6	2.5628	2.5570	0.23	0.5537	0.5506	0.56	10
0.8	9	4.2466	4.2447	0.05	0.6118	0.6090	0.46	29
0.9	14	9.3488	9.3413	0.08	0.6867	0.6796	1.04	105

Tabelle 5.2: Homogenes Fork/Join Modell aus zwei Stationen ($c_B^2 = 0.5$)

Die Tabelle 5.2 zeigt, daß auch im Fall $c_B^2 = 0.5$ die Approximationsfehler sowohl bzgl. des Erwartungswertes der Durchlaufzeitverteilung als auch bzgl. des Variationskoeffizienten sehr gering sind und zudem die in Tabelle 5.1 zusammengefaßten Werte unterschreiten. Aufgrund der geringeren Schwankungsbreite der Bedienzeiten sind zudem die U–Werte dieser Konfiguration geringer. Der Schwellwert der Auslastung, ab dem die Werte Schranken U deutlich stärker als linear anwachsen, verschiebt sich leicht auf Auslastungen $\rho \geq 0.8$. Die Laufzeit des Analyseverfahrens ist jedoch höher als im Fall negativ–exponentieller Bedienzeitverteilungen, da zur Darstellung der kleineren Variationskoeffizienten eine größere Anzahl Phasen benötigt wird.

Die Tabellen 5.3 und 5.4 zeigen die Resultate des homogenen Modells, die sich für die Variationskoeffizienten $c_B^2 = 2$ und $c_B^2 = 4$ der Bedienzeitverteilung ergeben. Ein deutliches Ansteigen der U–Werte läßt sich bereits für geringere Auslastungen ($\rho = 0.5$ bzw. $\rho = 0.4$) ausmachen. Konsequenterweise steigen die Laufzeiten des Analyseverfahrens für das Upper–Bound Modell mit wachsendem Variationskoeffizienten erheblich an. Die Approximationsfehler bleiben

jedoch weiterhin moderat.

		$E[D]$			c_D^2			
ρ	U	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
0.1	2	0.1850	0.1840	0.54	1.2160	1.2054	0.88	0
0.2	3	0.4347	0.4321	0.60	1.1760	1.1650	0.94	2
0.3	5	0.7735	0.7709	0.34	1.1153	1.1118	0.31	9
0.4	6	1.2504	1.2442	0.50	1.0521	1.0489	0.31	13
0.5	9	1.9356	1.9295	0.32	0.9769	0.9797	0.29	42
0.6	11	2.9973	2.9832	0.47	0.9044	0.9069	0.28	65
0.7	16	4.7834	4.7605	0.48	0.8284	0.8335	0.61	172
0.8	23	8.4041	8.3265	0.94	0.7581	0.7571	0.13	477
0.9	40	19.4277	19.0405	2.03	0.6958	0.6918	0.58	2394

Tabelle 5.3: Homogenes Fork/Join Modell aus zwei Stationen ($c_B^2 = 2$)

		$E[D]$			c_D^2			
ρ	U	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
0.1	2	0.2145	0.2124	0.99	2.3918	2.3671	1.04	1
0.2	5	0.5438	0.5418	0.37	2.1270	2.1190	0.38	10
0.3	7	1.0396	1.0345	0.49	1.8673	1.8592	0.44	19
0.4	10	1.7754	1.7662	0.52	1.6173	1.6098	0.47	58
0.5	14	2.8871	2.8713	0.55	1.3879	1.3821	0.42	133
0.6	20	4.6408	4.6123	0.62	1.1815	1.1762	0.45	326
0.7	27	7.6729	7.5955	1.02	1.0004	0.9915	0.9	576
0.8	43	13.8320	13.6030	1.68	0.8462	0.8332	1.56	1238
0.9	75	32.2461	31.6571	1.86	0.7236	0.7129	1.50	5598

Tabelle 5.4: Homogenes Fork/Join Modell aus zwei Stationen ($c_B^2 = 4$)

Die bisherigen Experimente haben gezeigt, daß die Analyse des Upper–Bound Modells zur Approximation eines homogenen Fork/Join–Netzes mit zwei Bedienern und negativ–exponentiell verteilten Zwischenankunftszeiten sehr gute Resultate hinsichtlich der ersten beiden Momente der Durchlaufzeitverteilung liefert. Im Falle großer Variationskoeffizienten der Bedienzeitverteilung muß jedoch mit erheblichen Laufzeiten gerechnet werden. Die Laufzeit läßt sich zweifelsohne durch die Vergrößerung des Schwellwertes ϵ und damit durch Verringerung der Schranken U reduzieren. Der Laufzeitgewinn muß jedoch erwartungsgemäß mit höheren Approximationsfehlern bezahlt werden. Diese Tatsache wird durch die Abbildungen 5.4 und 5.5 belegt.

Die Grafik 5.4 veranschaulicht das Anwachsen der Schranke U in Abhängigkeit des Schwellwertes ϵ für verschiedene Auslastungen ρ im Bereich $0.0005 \leq \epsilon \leq 0.05$. Für dasselbe Intervall zeigt die Abbildung 5.5 den mit steigendem ϵ höheren Approximationsfehler des Erwartungswertes der Durchlaufzeit.

Zum Abschluß der ersten Experimentreihe seien nochmals die Schranken U betrachtet. Die Tabellen 5.1 bis 5.4 haben die Abhängigkeit der Schranken U von der Auslastung der Bediener für unterschiedliche, jedoch feste Variationskoeffizienten der Bedienzeitverteilung auf-

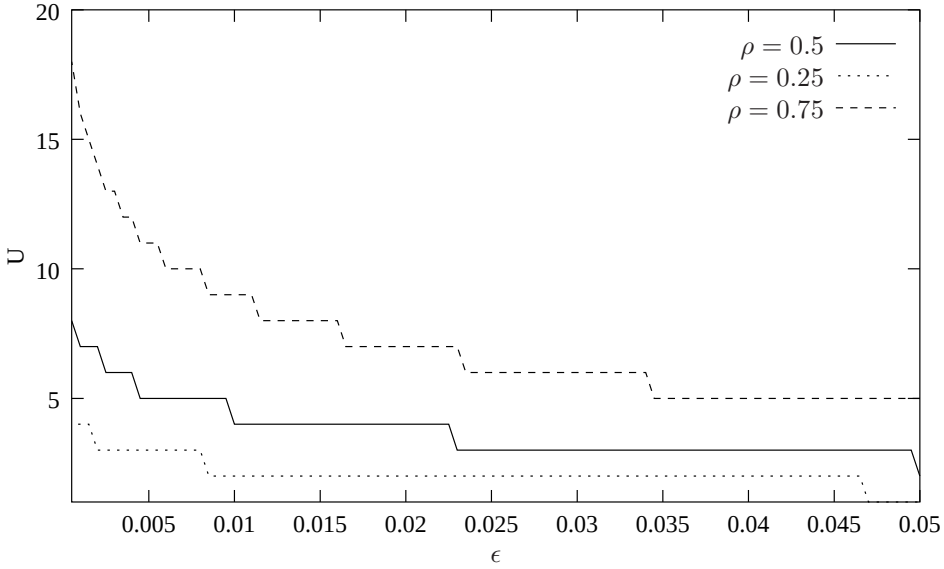

Abbildung 5.4: Abhängigkeit der U-Werte von ϵ

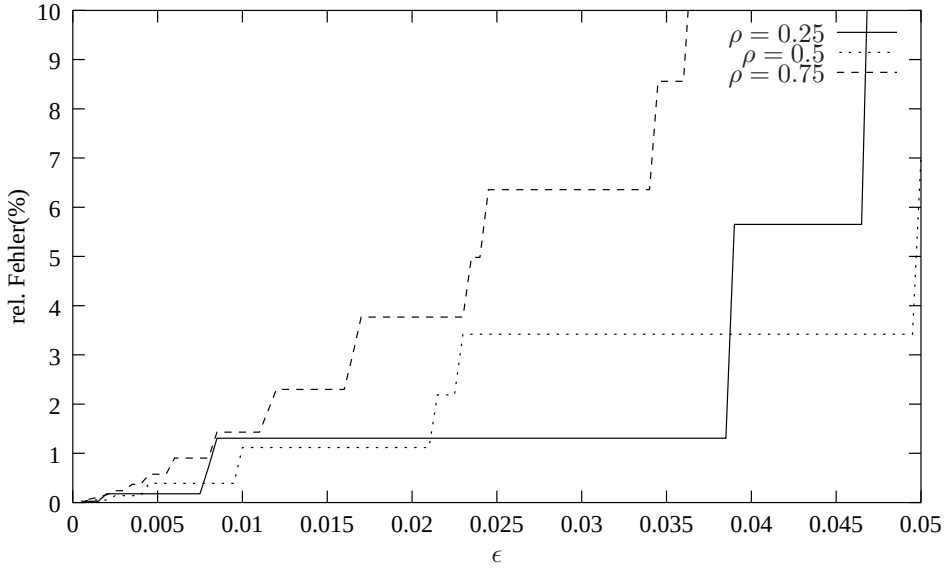

Abbildung 5.5: Approximationsgüte der Durchlaufzeit

gezeigt. In einem letzten Experiment, quer zu diesen Experimenten, soll gezeigt werden, wie sich die U–Werte bei fester Auslastung und variierendem Variationskoeffizienten verhalten. Diese Abhängigkeit wird in den Abbildungen 5.6 und 5.7 dargestellt. Für die Auslastungen $\rho = 0.25, \rho = 0.5$ (Abbildung 5.6) und $\rho = 0.75$ (Abbildung 5.7) ist das Anwachsen der Schranken U bei festem Schwellwert $\epsilon = 0.005$ und steigendem Variationskoeffizienten c_B^2 der Bedienzeitverteilung im Bereich $0.1 \leq c_B^2 \leq 20$ aufgetragen.

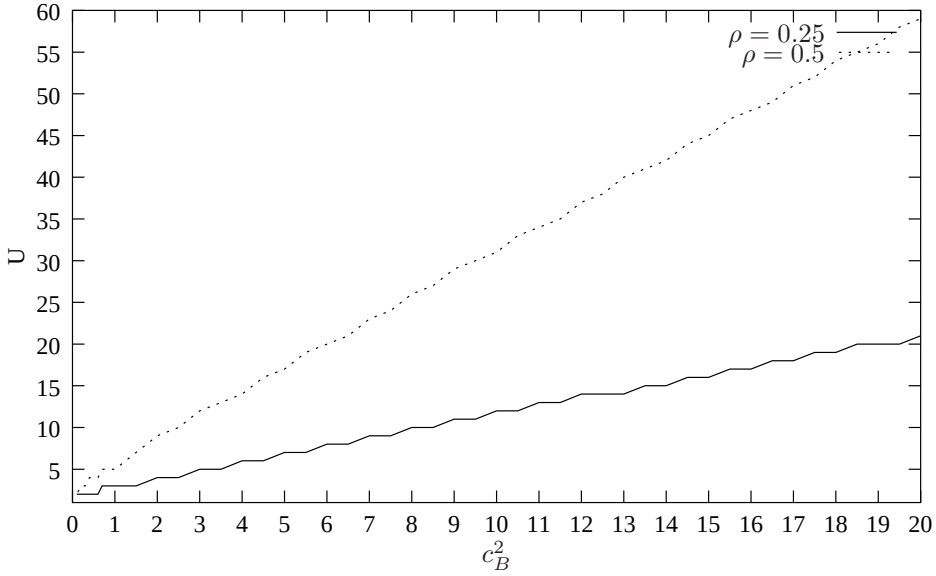

Abbildung 5.6: Abhängigkeit der $U_{i,j}$ von c_B^2 für $\lambda = 1, c_A^2 = 1, \mu = 2$

Die Abbildungen lassen die Vermutung zu, daß im Falle eines homogenen Fork/Join–Netzes und konstanter Auslastung der Bediener die Schranken U und der quadratische Variationskoeffizient der Bedienzeitverteilung in linearer Abhängigkeit stehen. Einer analytischen Untermauerung dieser Vermutung wurde im Rahmen dieser Arbeit nicht nachgegangen.

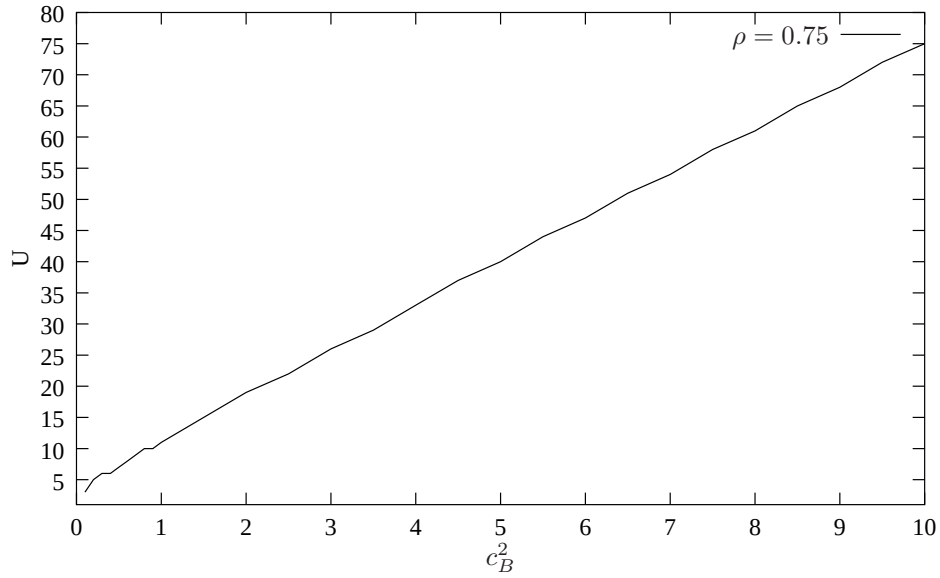

Abbildung 5.7: Abhängigkeit der $U_{i,j}$ von c_B^2 für $\lambda = 1, c_A^2 = 1, \mu = 2$

5.2.2 Experimentreihe 2: Inhomogenes Modell mit zwei parallelen Bedienern

In dieser zweiten Experimentreihe werden die Betrachtungen um den Fall eines inhomogenen Fork/Join–Netzes mit zwei Bedieneinrichtungen erweitert. Die Bedienzeitverteilungen der Aufträge an den Stationen besitzen unterschiedliche Raten, jedoch identische Variationskoeffizienten. Die Zwischenankunftszeiten bleiben unverändert negativ–exponentiell verteilt mit der Rate $\lambda = 1$. Die Schranken $U_{1,2}$ und $U_{2,1}$ wurden wiederum derartig gewählt, daß die Blockierungswahrscheinlichkeiten der Bediener gemäß Gleichung 5.57 den Schwellwert $\epsilon = 0.005$ unterschreiten. Im Gegensatz zur ersten Experimentserie werden sich die Schranken aufgrund der verschiedenen Auslastungen der beiden Bediener unterscheiden.

Die Resultate erster Untersuchungen sind in der Tabelle 5.5 zusammengefaßt. Die Bedienzeiten an beiden Bedienern besitzen negativ–exponentielle Verteilungen derart, daß sich für den ersten Bediener Auslastungen ρ_1 zwischen 0.1 und 0.9 einstellen. Die Raten der Bedienzeitverteilung an dem zweiten Bediener wurden um die Faktoren $1.5, 2$ und 3 größer gewählt, als die der ersten. Dementsprechend sind die Spalten der Tabelle 5.5 zu lesen. Die erste Spalte repräsentiert die Auslastung des ersten Bedieners, und die zweite Spalte gibt den Faktor an, um den sich die Rate μ_2 der zweiten Bedienzeitverteilung von der Rate μ_1 der ersten unterscheidet. Aufgrund der größeren Rate μ_2 und dem damit geringeren Erwartungswert wird die mittlere Anzahl an Aufträgen an der ersten Station die der zweiten Station übertreffen. Folglich ist bei festem $\epsilon = 0.005$ die Schranke $U_{1,2}$ größer als $U_{2,1}$. Die entsprechenden Werte sind in den Spalten 3 und 4 dargestellt. Die übrigen Spalten besitzen die bereits aus der ersten Experimentserie gewohnte Interpretation und vergleichen die Resultate des Analyseverfahrens für das Upper–Bound Modell mit der Simulation des primären Fork/Join–Netzes hinsichtlich des Erwartungswertes und des quadratischen Variationskoeffizienten der Durchlaufzeitverteilung. Schließlich erfaßt die letzte Spalte die Laufzeit, die das Analyseverfahren des Upper–Bound Modells zur Lösung des Modells benötigte.

Die Tabelle verdeutlicht, daß der Approximationsfehler wiederum sowohl hinsichtlich des Erwartungswertes als auch hinsichtlich des quadratischen Variationskoeffizienten in allen Experimenten sehr moderat ausfällt. Ein Vergleich mit der Tabelle 5.1 zeigt, daß die Laufzeit des Verfahrens etwa der des homogenen Falls bei gleicher Auslastung entspricht.

Deutlich interessanter ist jedoch die Beobachtung der Schranken $U_{1,2}$ und $U_{2,1}$. Die Betrachtung der Tabelle 5.5 legt die Vermutung nahe, daß bei fester Auslastung ρ_1 des ersten Bedieners die Summen $U_{1,2}+U_{2,1}$ der Schranken unabhängig von der Bedienrate des zweiten Bedieners etwa konstant bleiben. Ferner entspricht die Summe nahezu der Summe $U_{1,2}+U_{2,1} = 2U$ des homogenen Falls bei entsprechender Auslastung.

				$E[D]$			c_D^2			
ρ	μ_2	$U_{1,2}$	$U_{2,1}$	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
	$1.5\mu_1$	2	1	0.1396	0.1366	2.159	0.6246	0.6113	2.181	0
0.1	$2\mu_1$	2	1	0.1283	0.1263	1.59	0.6984	0.6914	0.9984	0
	$3\mu_1$	2	1	0.1194	0.118	1.188	0.8094	0.8115	0.2546	0
	$1.5\mu_1$	3	2	0.3055	0.3016	1.292	0.6376	0.6346	0.4799	0
0.2	$2\mu_1$	3	1	0.2855	0.2794	2.215	0.7236	0.7197	0.5404	0
	$3\mu_1$	3	1	0.2663	0.2622	1.571	0.8321	0.8339	0.2148	0
	$1.5\mu_1$	4	2	0.514	0.5065	1.476	0.6613	0.6585	0.4207	0
0.3	$2\mu_1$	4	2	0.4758	0.4703	1.157	0.7501	0.7535	0.4549	0
	$3\mu_1$	4	1	0.4529	0.4451	1.764	0.8548	0.8617	0.8071	0
	$1.5\mu_1$	5	3	0.7793	0.7715	1.016	0.6857	0.6846	0.1654	1
0.4	$2\mu_1$	5	2	0.7278	0.7205	1.021	0.7823	0.7844	0.2796	1
	$3\mu_1$	5	1	0.699	0.6888	1.485	0.8774	0.8871	1.088	1
	$1.5\mu_1$	6	3	1.142	1.135	0.6194	0.7193	0.7231	0.5252	1
0.5	$2\mu_1$	7	2	1.074	1.064	0.8878	0.8176	0.8186	0.1182	1
	$3\mu_1$	7	2	1.029	1.024	0.4933	0.9101	0.9137	0.3987	1
	$1.5\mu_1$	8	3	1.671	1.655	0.9292	0.7605	0.76	0.06188	1
0.6	$2\mu_1$	9	2	1.585	1.575	0.6769	0.8557	0.858	0.2733	2
	$3\mu_1$	9	2	1.531	1.529	0.1212	0.9349	0.935	0.007005	2
	$1.5\mu_1$	12	4	2.517	2.506	0.4036	0.8116	0.8131	0.1755	4
0.7	$2\mu_1$	13	3	2.412	2.406	0.2531	0.9009	0.9016	0.07976	4
	$3\mu_1$	14	2	2.364	2.36	0.1533	0.9579	0.959	0.1073	4
	$1.5\mu_1$	19	4	4.195	4.168	0.6663	0.8748	0.8799	0.5735	10
0.8	$2\mu_1$	21	3	4.075	4.061	0.3648	0.943	0.9464	0.3517	11
	$3\mu_1$	22	2	4.029	4.019	0.2494	0.9777	0.9811	0.3395	11
	$1.5\mu_1$	41	4	9.199	9.167	0.3426	0.9439	0.952	0.8552	52
0.9	$2\mu_1$	44	3	9.114	9.081	0.3634	0.9793	0.9811	0.175	56
	$3\mu_1$	46	2	9.078	9.052	0.2872	0.9926	0.9923	0.02941	60

Tabelle 5.5: Heterogenes Upper–Bound Modell aus zwei Stationen ($\lambda = 1, c_A^2 = 1, c_B^2 = 1$)

In zwei weiteren Experimenten wurden die quadratischen Variationskoeffizienten $c^2_{B_1}$ und $c^2_{B_2}$ beider Bedienzeitverteilungen modifiziert und auf die Werte $c^2_{B_1} = c^2_{B_2} = 0.5$ und $c^2_{B_1} = c^2_{B_2} = 2$ eingestellt. Die zugehörigen Resultate sind in den Tabellen 5.6 und 5.7 dargestellt. Wiederum scheinen die Summen $U_{1,2} + U_{2,1}$ bei fester Auslastung des ersten Bedieners nahezu konstant zu sein. Im Fall $c^2_{B_1} = c^2_{B_2} = 0.5$ entspricht diese zudem etwa der Summe der Schranken des homogenen Falls bei gleicher Auslastung. Lediglich bei hoher Auslastung $\rho_1 = 0.9$ bildet der Wert des homogenen Modells eine untere Schranke für das inhomogene Modell.

Ein ähnliches Verhalten läßt sich für das inhomogene Modell mit den quadratischen Variationskoeffizienten $c^2_{B_1} = c^2_{B_2} = 2$ beobachten. In diesem Fall bilden die Schranken des homogenen Modells eine obere Schranke für das inhomogene Modell, jedoch bereits ab Auslastungen von 0.5. Diese Auslastungsgrenzen 0.9 im Fall $c^2_{B_1} = c^2_{B_2} = 0.5$ und 0.5 im Fall $c^2_{B_1} = c^2_{B_2} = 2$ sind genau die Grenzen, für die die Werte U des homogenen Modells deutlich stärker als linear anwachsen.

				$E[D]$			c^2_D			
ρ	μ_2	$U_{1,2}$	$U_{2,1}$	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
	$1.5\mu_1$	2	1	0.1262	0.1245	1.329	0.3545	0.3416	3.78	0
0.1	$2\mu_1$	2	1	0.1177	0.1164	1.149	0.4024	0.3919	2.675	0
	$3\mu_1$	2	1	0.1119	0.1108	0.9482	0.4637	0.4557	1.754	0
	$1.5\mu_1$	2	2	0.2716	0.2696	0.7377	0.3828	0.3753	1.999	1
0.2	$2\mu_1$	2	1	0.2567	0.2519	1.876	0.4402	0.4325	1.767	0
	$3\mu_1$	2	1	0.2445	0.2418	1.11	0.4998	0.4963	0.7067	0
	$1.5\mu_1$	3	2	0.4471	0.4411	1.364	0.421	0.4118	2.249	1
0.3	$2\mu_1$	3	1	0.4257	0.4164	2.218	0.482	0.4766	1.122	1
	$3\mu_1$	3	1	0.4066	0.401	1.398	0.5401	0.5348	0.9875	1
	$1.5\mu_1$	4	2	0.6669	0.6586	1.251	0.4673	0.4529	3.178	2
0.4	$2\mu_1$	4	2	0.6305	0.6235	1.126	0.5298	0.517	2.482	2
	$3\mu_1$	4	1	0.6132	0.6026	1.76	0.5854	0.5763	1.562	1
	$1.5\mu_1$	5	2	0.9576	0.9402	1.851	0.523	0.5047	3.638	3
0.5	$2\mu_1$	5	2	0.9101	0.8937	1.845	0.5874	0.5662	3.743	3
	$3\mu_1$	5	1	0.8912	0.8723	2.158	0.636	0.6181	2.908	2
	$1.5\mu_1$	6	2	1.373	1.347	1.897	0.59	0.5646	4.492	4
0.6	$2\mu_1$	7	2	1.313	1.295	1.376	0.6536	0.6285	4.004	5
	$3\mu_1$	7	1	1.294	1.273	1.662	0.6929	0.6771	2.329	4
	$1.5\mu_1$	9	3	2.02	1.991	1.451	0.6713	0.6729	0.2394	10
0.7	$2\mu_1$	10	2	1.964	1.94	1.238	0.7291	0.7313	0.3092	10
	$3\mu_1$	11	1	1.947	1.912	1.826	0.7567	0.7637	0.9245	10
	$1.5\mu_1$	14	3	3.294	3.244	1.529	0.7712	0.7597	1.516	26
0.8	$2\mu_1$	16	2	3.237	3.193	1.37	0.8134	0.8156	0.2638	31
	$3\mu_1$	17	1	3.225	3.177	1.509	0.8284	0.8345	0.7299	32
	$1.5\mu_1$	30	3	7.363	7.164	2.78	0.8856	0.8971	1.284	139
0.9	$2\mu_1$	33	2	7.266	7.114	2.13	0.9051	0.9155	1.14	162
	$3\mu_1$	35	1	7.194	7.097	1.37	0.9093	0.9202	1.19	180

Tabelle 5.6: Heterogenes Upper–Bound Modell aus zwei Stationen ($\lambda = 1, c^2_A = 1, c^2_B = 0.5$)

				$E[D]$			c_D^2			
ρ	μ_2	$U_{1,2}$	$U_{2,1}$	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
0.1	$1.5\mu_1$	2	1	0.1543	0.1505	2.525	1.302	1.294	0.6819	0
	$2\mu_1$	2	1	0.1404	0.1376	2.002	1.417	1.413	0.3204	0
	$3\mu_1$	2	1	0.1289	0.127	1.461	1.6	1.599	0.05911	0
0.2	$1.5\mu_1$	3	2	0.3533	0.3457	2.184	1.282	1.267	1.162	2
	$2\mu_1$	3	2	0.322	0.3165	1.76	1.414	1.394	1.422	2
	$3\mu_1$	3	1	0.3004	0.2928	2.591	1.594	1.576	1.145	1
0.3	$1.5\mu_1$	5	3	0.6171	0.6075	1.57	1.249	1.238	0.9434	5
	$2\mu_1$	5	2	0.5662	0.5559	1.847	1.396	1.38	1.115	3
	$3\mu_1$	5	2	0.5267	0.5181	1.651	1.578	1.564	0.9299	4
0.4	$1.5\mu_1$	7	4	0.9754	0.957	1.93	1.212	1.21	0.1718	11
	$2\mu_1$	7	3	0.8964	0.8764	2.281	1.367	1.36	0.4897	9
	$3\mu_1$	7	2	0.8441	0.8291	1.814	1.536	1.535	0.02537	7
0.5	$1.5\mu_1$	9	5	1.481	1.466	0.9959	1.172	1.185	1.174	21
	$2\mu_1$	10	3	1.373	1.357	1.138	1.328	1.35	1.641	18
	$3\mu_1$	10	2	1.304	1.296	0.637	1.477	1.509	2.133	15
0.6	$1.5\mu_1$	13	5	2.244	2.227	0.7536	1.133	1.142	0.7842	40
	$2\mu_1$	14	4	2.089	2.071	0.8669	1.285	1.3	1.169	42
	$3\mu_1$	14	2	2.013	1.983	1.522	1.403	1.424	1.464	32
0.7	$1.5\mu_1$	19	6	3.497	3.472	0.7329	1.098	1.122	2.155	99
	$2\mu_1$	20	4	3.308	3.27	1.152	1.232	1.254	1.777	94
	$3\mu_1$	21	2	3.222	3.175	1.481	1.318	1.344	1.909	85
0.8	$1.5\mu_1$	30	7	5.984	5.873	1.876	1.072	1.072	0.00789	271
	$2\mu_1$	32	4	5.772	5.669	1.833	1.17	1.173	0.2521	252
	$3\mu_1$	33	2	5.682	5.58	1.835	1.222	1.224	0.1759	242
0.9	$1.5\mu_1$	62	7	13.47	13.34	0.9772	1.046	1.078	2.957	1783
	$2\mu_1$	67	4	13.24	13.2	0.2551	1.095	1.125	2.706	1698
	$3\mu_1$	70	2	13.15	13.14	0.07361	1.115	1.147	2.717	1604

Tabelle 5.7: Heterogenes Upper–Bound Modell aus zwei Stationen ($\lambda = 1, c_A^2 = 1, c_B^2 = 2$)

Diese Eigenschaften der Schranken U wirken sich unmittelbar auf die Laufzeiten des Analyseverfahrens aus. Da für Variationskoeffizienten $c_{B_1} = c_{B_2} < 1$ die Summe der $U_{i,j}$ des homogenen Modells eine untere Grenze für die Summe der $U_{i,j}$ des inhomogenen Modells bilden, ist auch die Laufzeit des Analyseverfahrens auf dem homogenen Modell eine untere Grenze für die Laufzeit auf dem inhomogenen Modell. Mit der gleichen Argumentation ist die Laufzeit der Analyse des homogenen Modells im Fall $c_{B_1} = c_{B_2} > 1$ eine obere Grenze für die Laufzeit auf dem inhomogenen Modell bei gleicher Auslastung.

Zusammenfassend läßt sich festhalten, daß die Durchlaufzeit eines inhomogenen Fork/Join–Netzes, dessen Bedienzeiten identische Variationskoeffizienten besitzen, deutlich von dem Bediener mit höherer Auslastung dominiert wird. Der Approximationsfehler hinsichtlich der ersten beiden Momente der Durchlaufzeit des Upper–Bound Modells gegenüber dem primären Modell ist sehr gering. Das Problem hoher Laufzeiten bleibt weiterhin präsent, wird jedoch gegenüber dem homogenen Modell leicht gemildert, da insbesondere im Fall großer Variationskoeffizienten mit geringeren Laufzeiten als im homogenen Fall zu rechnen ist.

5.2.3 Experimentreihe 3: Erweitertes inhomogenes Modell mit zwei parallelen Bedienern

Die Ausführungen der zweiten Experimentreihe haben belegt, daß die Durchlaufzeit eines inhomogenen Fork/Join–Netzes mit zwei parallelen Bedieneinrichtungen und identischen Variationskoeffizienten der Bedienzeitverteilungen von dem höher ausgelasteten Bediener dominiert wird. Die Verringerung des Variationskoeffizienten der Bedienzeitverteilung mit höherer Rate verstärkt diesen Effekt zusätzlich. In dieser dritten Experimentreihe wird daher untersucht, wie sich die gleichzeitige Erhöhung der Raten und des Variationskoeffizienten der zweiten Bedienzeitverteilung auswirkt.

Wie in den vorangegangenen Experimenten wird die Rate μ_1 derart eingestellt, daß sich für den ersten Bediener Auslastungen zwischen 0.1 und 0.9 einstellen. Der Variationskoeffizient dieser ersten Bedienzeitverteilung wird mit $c^2_{B_1} = 0.5$ fest gewählt. Die Bedienrate μ_2 unterscheidet sich von μ_1 um die Faktoren $1.5, 2$ und 3. Der Variationskoeffizient c_{B_2} wird fest auf $c^2_{B_2} = 4$ gesetzt. Die Zwischenankunftszeiten sind wiederum negativ–exponentiell mit der Rate $\lambda = 1$ verteilt.

Die Tabelle 5.8 faßt die Resultate dieser dritten Experimentreihe zusammen. Durch den Vergleich der Werte $U_{1,2}$ und $U_{2,1}$ zeigt sich, daß sich die Dominanz des ersten Bedieners erst bei sehr hohen Auslastungen bemerkbar macht. Die recht große Schwankungsbreite der zweiten Bedienzeitverteilung resultiert in hohen Werten $U_{2,1}$.

Interessant ist die Beobachtung, daß sich die Werte $U_{1,2}$ dieses Modells nicht von den Werten $U_{1,2}$ des inhomogenen Modells mit identischen Variationskoeffizienten $c^2_{B_1} = c^2_{B_2} = 0.5$ unterscheiden (vgl. dazu Tab. 5.6). Diese Beobachtung legt die Vermutung nahe, daß die Größe der Schranke $U_{1,2}$ wohl von der Rate μ_2 der Bedienzeitverteilung von Aufträgen an der zweiten Bedieneinrichtung abhängt, nicht jedoch vom Variationskoeffizienten. Diese Vermutung konnte anhand weiterer Experimente bestätigt werden.

Der Blick auf den Approximationsfehler dieser Experimente läßt wiederum sehr kleine Abweichungen erkennen. Im Fall hoher Auslastungen des ersten Bedieners ist der Fehler offensichtlich signifikant höher, als in den bisherigen Experimenten, jedoch prozentual immer noch recht gering.

				$E[D]$			c_D^2			
ρ_1	μ_2	$U_{1,2}$	$U_{2,1}$	UB	Sim	Δ (%)	UB	Sim	Δ (%)	CPU (Sek.)
	μ_1	2	2	0.1848	0.1819	1.57	1.7246	1.6877	2.19	0
	$1.5\mu_1$	2	2	0.1468	0.1459	0.6575	1.101	1.081	1.799	0
0.1	$2\mu_1$	2	1	0.1335	0.1309	2.009	0.8322	0.7863	5.841	0
	$3\mu_1$	2	2	0.1199	0.1191	0.6986	0.5919	0.5731	3.269	0
	μ_1	2	5	0.4427	0.436964	1.31	1.7684	1.7558	0.72	3
	$1.5\mu_1$	2	3	0.3321	0.3276	1.348	1.131	1.114	1.6	1
0.2	$2\mu_1$	2	2	0.2938	0.2893	1.573	0.8337	0.805	3.568	1
	$3\mu_1$	2	2	0.2633	0.2602	1.171	0.6105	0.5901	3.46	1
	μ_1	3	7	0.80608	0.798518	0.947	1.744	1.7426	0.008	7
	$1.5\mu_1$	3	5	0.5654	0.5578	1.363	1.122	1.111	1.033	4
0.3	$2\mu_1$	3	3	0.4912	0.4836	1.555	0.8244	0.8049	2.426	2
	$3\mu_1$	3	2	0.4392	0.4324	1.582	0.6279	0.6054	3.724	2
	μ_1	3	11	1.3293	1.319846	0.716	1.673	1.66	0.78	16
	$1.5\mu_1$	4	6	0.8684	0.8571	1.313	1.093	1.075	1.683	7
0.4	$2\mu_1$	4	4	0.7409	0.7282	1.743	0.8075	0.79	2.213	4
	$3\mu_1$	4	3	0.6575	0.6512	0.9646	0.6349	0.63	0.7803	3
	μ_1	4	15	2.1034	2.076455	1.298	1.569	1.555	0.9	38
	$1.5\mu_1$	5	8	1.268	1.26	0.6612	1.038	1.03	0.7544	16
0.5	$2\mu_1$	5	5	1.069	1.053	1.543	0.7837	0.7772	0.8363	7
	$3\mu_1$	5	3	0.9527	0.9411	1.225	0.6576	0.6553	0.3473	4
	μ_1	5	20	3.3208	3.276409	1.355	1.444	1.437	0.49	87
	$1.5\mu_1$	7	9	1.825	1.805	1.139	0.964	0.9385	2.723	24
0.6	$2\mu_1$	7	6	1.527	1.513	0.9095	0.7569	0.7574	0.07446	14
	$3\mu_1$	7	4	1.365	1.359	0.4172	0.6815	0.6994	2.569	10
	μ_1	7	28	5.4162	5.339	1.45	1.309	1.284	1.95	206
	$1.5\mu_1$	9	11	2.656	2.621	1.344	0.8725	0.8508	2.547	51
0.7	$2\mu_1$	10	7	2.234	2.216	0.7933	0.7387	0.734	0.6361	28
	$3\mu_1$	11	4	2.034	2.017	0.829	0.7263	0.7328	0.8915	22
	μ_1	10	41	9.719	9.534	1.95	1.17	1.135	3.08	640
	$1.5\mu_1$	15	13	4.118	4.034	2.091	0.7792	0.7752	0.5045	118
0.8	$2\mu_1$	16	7	3.581	3.497	2.397	0.7536	0.7677	1.838	67
	$3\mu_1$	17	4	3.33	3.282	1.47	0.7904	0.817	3.251	62
	μ_1	17	74	22.918	22.614	1.35	1.033	1.0477	1.4	2067
	$1.5\mu_1$	31	13	8.121	7.672	5.858	0.7567	0.7522	0.6051	583
0.9	$2\mu_1$	34	7	7.451	7.17	3.922	0.8246	0.8401	1.84	329
	$3\mu_1$	35	4	7.132	6.952	2.587	0.8791	0.8839	0.5347	304

Tabelle 5.8: Inhomogenes Modell ($\lambda = 1, c_A^2 = 1, c_{B_1}^2 = 0.5, c_{B_2}^2 = 4$)

5.2.4 Experimentreihe 4: Homogenes Modell mit drei parallelen Bedienern

In einem letzten Experiment wird ein Modell mit drei parallelen Bedienern betrachtet. Die identisch negativ–exponentiell verteilten Bedienzeiten der drei Bediener werden derart eingestellt, daß sich für die Bediener unter dem ebenfalls negativ–exponentiell verteilten Ankunftsprozeß mit der Rate $\lambda = 1$ Auslastungen zwischen 0.1 und 0.9 ergeben. Die Resultate dieser Experimentreihe sind in der Tabelle 5.9 dargestellt.

	$E[D]$			c_D^2				
ρ_1	UB	Sim	Δ (%)	UB	Sim	Δ (%)	U	Zustände
0.1	0.201	0.199	1.00	0.421	0.414	1.69	2	19
0.2	0.448	0.442	1.36	0.447	0.423	5.67	2	19
0.3	0.753	0.752	0.2	0.461	0.441	4.54	3	37
0.4	1.155	1.153	0.2	0.484	0.461	4.99	4	61
0.5	1.731	1.728	0.2	0.502	0.489	2.66	5	91
0.6	2.583	2.543	1.57	0.511	0.504	1.39	7	169
0.7	3.941	3.875	1.7	0.520	0.499	4.00	9	271
0.8	6.634	6.538	1.47	0.527	0.509	3.54	13	547
0.9	15.037	14.593	3.04	0.531	0.539	1.48	23	1657

Tabelle 5.9: Inhomogenes Modell ($\lambda = 1, c_A^2 = 1, c_B^2 = 1.0$)

Es zeigt sich, daß auch im Fall von mehr als zwei parallelen Bedienern die Approximationsgüte der Analyseresultate sehr hoch ist. Nachteilig ist jedoch der mit steigender Anzahl paralleler Bediener stark anwachsende Berechnungsaufwand sowie der enorm anwachsende Speicherplatzbedarf. Um diese Tatsache zu belegen, sei für den vorliegenden Fall die Anzahl der Makrozustände in den Partitionen $\tilde{Z}_K$ (vgl. Gleichung (5.9) in Abschnitt 5) betrachtet. Diese Anzahl sowie die Anzahl der Mikrozustände je Makrozustand bestimmt die Größe der zur Analyse des Upper–Bound Modells benötigten Matrizen. Aufgrund der jeweils identisch verteilten Bedienzeiten in dem betrachteten Modell gilt für die (identischen) Schranken $U_{i,j}, i \neq j = 1, \ldots, 3$ die Beziehung $U_{i,j} = U_{j,i}$. Es ist leicht einzusehen, daß damit $\tilde{Z}_k$ für alle $k \geq 0$ $\prod_{i=1}^{3} \prod_{j=i+1}^{3} (U_{i,j} + 1) - \prod_{i=1}^{3} \prod_{j=i+1}^{3} U_{i,j}$ Zustände enthält. Aufgrund der negativ–exponentiell verteilten Ankunfts– und Bedienprozesse enthält jeder Makrozustand genau einen Mikrozustand. Die Schranken $U = U_{i,j}$ sowie die Größen der Makrozustandsräume $\tilde{Z}_k$ sind in den letzten beiden Spalten der Tabelle 5.9 angegeben. Im Fall $\mathcal{H}$ verteilter Ankunfts– und Bedienprozesse entspricht die Anzahl der Mikrozustände je Makrozustand dem Produkt der Phasenanzahlen des Ankunftsprozesse und der Bedienprozesse, so daß in diesem Fall die Effizienz des Verfahren nochmals drastisch abnimmt. Auf diese Problematik wird nochmals in Abschnitt 6.4 sowie in einem Anwendungsbeispiel in Abschnitt 7.1 eingegangen. Ferner wird ein Lösungsansatz zur Effizienzsteigerung aufgezeigt, der statt eines Fork/Join–Netzes mit $M > 2$ parallelen Bedienern $M - 1$ Fork/Join–Netze aus 2 parallelen Bedienern betrachtet. Dieser Ansatz läßt sich die im zweiten Teil der Arbeit vorgestellte Aggregierungstechnik erreichen. Dabei wird sich jedoch im allgemeinen ein höherer Approximationsfehler einstellen, was in Abschnitt 7.1 deutlich wird.

5.3 Das Upper–Bound Modell in einem Warteschlangenkontext

Die Untersuchungen der vorangegangenen Abschnitte haben verdeutlicht, daß sich das Analyseverfahren für das Upper–Bound Modell zur Approximation einer von ihrer Umgebung losgelösten Fork/Join–Station in vielen Fällen sehr gut eignet. Hinsichtlich der ersten beiden Momente der Durchlaufzeitverteilung ergaben sich lediglich geringe Approximationsfehler. Konsequenterweise resultiert daraus die Überlegung, ob diese erfreulichen Ergebnisse im Kontext einer Umgebung erhalten bleiben. Konkret ist zu klären, ob sich dieses Analyseverfahren in der Kombination mit dem Dekompositionsverfahren nach Kühn/Whitt bewährt. Von besonderem Interesse ist zudem die Fragestellung, inwieweit sich die Eigenschaft des Upper–Bound Modells, die mittlere Durchlaufzeit der primären Fork/Join–Station nach oben abzuschätzen, auf die mittlere Durchlaufzeit von Fork/Join–Warteschlangennetzen übertragen läßt. Beide Aspekte sind eng mit der Ausprägung der betrachteten Modelltypen verbunden. Im folgenden werden Fork/Join–Warteschlangennetze betrachtet, die die in Abschnitt 3.2 behandelten PH/PH/1-∞ Stationen sowie Fork/Join Stationen enthalten.

Sei also zunächst die Erwartungshaltung dargelegt, die der Analyse derartiger Netze anhand des Dekompositionsverfahrens hinsichtlich obiger Fragestellungen entgegengebracht werden muß. Im Vordergrund dieser Überlegungen stehen wiederum die Auswirkungen auf die Durchlaufzeitverteilungen sowohl der isolierten Subnetze als auch des Gesamtnetzes. Welche Faktoren kommen also in Betracht, obige Fragestellungen negativ zu beantworten, d.h. welche Aspekte beeinflussen die Analyse der hier im Fokus stehenden Fork/Join–Warteschlangennetze. Die Faktoren sind einerseits in den Prinzipien des Dekompositionsverfahrens und andererseits in der Approximationsgüte der Analyse der beiden Stationstypen zu suchen.

Die Prinzipien des Dekompositionsverfahrens sind:

1. Die Zerlegung eines Warteschlangennetzes in disjunkte Subnetze, die isoliert voneinander analysiert werden.

2. Die Approximation der Schnittstellen der Subnetze, d.h. der Ankunfts– und Abgangsprozesse durch stationäre Erneuerungsprozesse.

3. Die ausschließliche Betrachtung der Schnittstellenprozesse und der Bedienprozesse hinsichtlich der ersten beiden Momente.

Der erste Aspekt hat unmittelbar Konsequenzen für das zweite Moment bzw. den Variationskoeffizienten der Gesamtdurchlaufzeitverteilung. Die isolierte Analyse der Subnetze und die daraus abgeleitete Komposition netzweiter Resultate birgt einen Informationsverlust, da diese Vorgehensweise Korrelationen unter den Durchlaufzeiten der Subnetze unberücksichtigt läßt. Diese inhärente Problematik des Dekompositionsverfahrens läßt folglich nur sehr vage Aussagen über das zweite Moment der Gesamtdurchlaufzeit zu. Hinsichtlich des Erwartungswertes bleibt dies jedoch ohne Folgen, so daß dieser im weiteren Verlauf ins Zentrum der Überlegungen rückt.

Das zweite Prinzip, also die Approximation der Ankunfts– und Abgangsprozesse der Stationen durch stationäre Erneuerungsprozesse, wird unterschiedliche Auswirkungen haben und sowohl von der Struktur des betrachteten Fork/Join–Warteschlangennetzes also auch von den Auslastungen der Stationen abhängen. Dies wird zunächst am Beispiel der PH/PH/1/∞–FCFS Station deutlich. Gemäß des Resultats von Marshall [68] (vgl. auch Abschnitt 2.2.2) ist im Falle einer GI/G/1/∞-FCFS Station die folgende Formel 5.58 eine gute Approximation für den Variationskoeffizienten c_D des Abgangsprozesses:

$$c_D^2 = c_A^2 + \rho^2(c_B^2 - c_A^2) \tag{5.58}$$

Die Formel besagt, daß c_D^2 bei niedriger Auslastung ρ im wesentlichen von dem Variationskoeffizienten c_A des Ankunftsprozesses bestimmt wird. Mit steigender Auslastung nimmt jedoch der Einfluß von c_A ab, und der Einfluß des Variationskoeffizienten c_B der Bedienzeitverteilung nimmt zu. Ließe sich diese Formel verallgemeinern, so bedeutete dies, daß der Abgangsprozeß der niedrig ausgelasteten GI/G/1/∞ Station im wesentlichen die Eigenschaften des Ankunftsprozesses aufwiese, und der Abgangsprozeß der hoch ausgelasteten GI/G/1/∞ Station hauptsächlich durch den Bedienprozeß beeinflußt würde. Übertragen auf den vorliegenden Fall der PH/PH/1/∞ Station reicht folglich ein niedrig ausgelasteter Bediener Approximationsfehler aufgrund der fälschlichen Annahme eines Erneuerungs–Ankunftsprozesses an den Abgangsprozeß und damit an seine Nachfolger weiter. Eine hoch ausgelastete PH/PH/1/∞ Station hingegen überträgt die Eigenschaften der unabhängig identisch verteilten Bedienzeiten auf den Abgangsprozeß. In diesem Fall ist also damit zu rechnen, daß Approximationsfehler aufgrund der fälschlichen Annahme eines Erneuerungs–Ankunftsprozesses für die nachfolgenden Stationen gefiltert werden.

Etwas differenzierter verhält sich die Situation für Fork/Join–Stationen. In Abschnitt 5.1.3 wurde verdeutlicht, daß die Fork/Join Station zur Berechnung des Abgangsprozesses des Upper–Bound Modells als spezielle Single–Server–Station aufgefaßt werden kann. Die Bedienzeiten dieser Single–Server Station sind jedoch keineswegs unabhängig verteilt. Es ist zu vermuten, daß auch der Abgangsprozeß der Fork/Join–Station auf ähnliche Weise wie im Falle der PH/PH/1/∞ Station durch den Ankunftsprozeß und die Bedienprozesse der parallelen Stationen beeinflußt wird. Folglich ist insbesondere die fälschliche Approximation des Abgangsprozesses einer Fork/Join–Station mit hoch ausgelasteten parallelen Bedienern durch stationäre Erneuerungsprozesse eine zusätzliche Fehlerquelle.

Zudem wird in zyklischen Netzen die Mißachtung von Korrelation in den Ankunftsprozessen hinsichtlich der Analyseresultate von Bedeutung sein.

Zur Bewertung des dritten Aspektes wird zunächst eine Behauptung/Beobachtung von Krämer/Langenbach–Belz [53] (vgl. Abschnitt 2.2.2) reflektiert. Demnach hängt der Erwartungswert einer GI/G/1-∞ Station mit FCFS Bedienung hauptsächlich von den ersten beiden Momenten des Ankunfts– und des Bedienprozesses ab. Folglich gilt diese Aussage auch für die hier betrachteten PH/PH/1/∞ Stationen. Da sich die Fork/Join–Stationen aus PH/PH/1/∞ Stationen zusammensetzt, ist zu vermuten, daß obige Behauptung auch in diesem Fall Gültigkeit behält. Der dritte Aspekt gibt somit zumindest hinsichtlich des Erwartungswertes der Durchlaufzeitverteilungen keinen Anlaß zur Sorge um die Analyseresultate des Dekompositi-

onsverfahrens.

Zusammenfassend resultieren aus dem Dekompositionsverfahren im wesentlichen die folgenden drei Fehlerquellen:

1. Die fehlende Berücksichtigung von Korrelationen unter den Durchlaufzeitverteilungen der isolierten Stationen zur Ermittlung der Gesamtdurchlaufzeitverteilung.
2. Die Approximation des Abgangsprozesses einer Fork/Join Station mit hoch ausgelasteten parallelen Bedienern durch einen stationären Erneuerungsprozeß.
3. Die fehlende Beachtung von Korrelationen der Ankunftsprozesse in zyklischen Netzen.

Seien im weiteren die Faktoren betrachtet, die sich aus der Approximationsgüte der Resultate der isolierten Stationen ergeben. Die Analyse der in Abschnitt 3.2 behandelten PH/PH/1-∞ Station ist exakt und kann somit unter dem Gesichtspunkt Approximationsgüte vernachlässigt werden. Über die Qualität der Resultate des Upper–Bound Modells hinsichtlich der Momente der Durchlaufzeitverteilung wurde im Abschnitt 5.2 ein sehr positives Fazit gezogen. Folglich bleibt die Beurteilung der Schnittstelle des Upper–Bound Modells, insbesondere die Bewertung der Approximationsgüte des zweiten Moments des Abgangsprozesses übrig (das erste Moment ist exakt und stimmt mit dem ersten Moment des Ankunftsprozesses überein). Der hierdurch induzierte Fehler ist jedoch sehr gering. Diese Behauptung wird anhand der Resultate der Experimentserie 3 in Abschnitt 5.2.3 (vgl. Tabelle 5.8) belegt. In dieser Experimentserie resultierten hinsichtlich der Momente der Durchlaufzeitverteilung für die isolierte Fork/Join-Station sehr geringe Approximationsfehler, die jedoch im Vergleich mit den übrigen Serien deutlicher ausfielen. Die Tabelle 5.10 vergleicht die zugehörigen Ergebnisse hinsichtlich der quadratischen Variationskoeffizienten der Zwischenabgangszeiten, die sich einerseits aus dem Analyseverfahren für das Upper–Bound Modell und andererseits aus der Simulation ergaben. Die relative prozentuale Abweichung Δ verdeutlicht sehr gute Resultate des Analyseverfahrens für das Upper–Bound Modell. Unter der Berücksichtigung der oben erläuterten Approximationen des Dekompositionsverfahrens ist der Fehler des Variationskoeffizienten des Abgangsprozesses als vernachlässigbar anzusehen.

Mit dem Hintergrund dieser Überlegungen sind in azyklischen Fork/Join–Warteschlangennetzen sehr gute Analyseresultate hinsichtlich der Erwartungswerte der Durchlaufzeitverteilungen der isolierten Stationen als auch des gesamten Netzes zu erwarten. Die einzige wesentliche Fehlerquelle liegt in der Approximation des Abgangsprozesses der Fork/Join–Station durch einen stationären Erneuerungsprozeß. Hinsichtlicht des zweiten Moments der Gesamtdurchlaufzeitverteilung kommt weiterhin die fehlende Beachtung der Korrelationen unter den Durchlaufzeiten der isolierten Stationen hinzu.

In zyklischen Netzen werden sich aufgrund von Korrelationen in den Ankunftsprozessen im allgemeinen größere Fehler ergeben. Diese Fehlerquelle wird zudem dominierend sein. In einem konkreten Kontext ist denkbar, daß die Eigenschaft des Upper–Bound Modells, den Erwartungswert der Durchlaufzeit nach oben abzuschätzen, nicht erhalten bleibt. Dieser Fall tritt insbesondere dann ein, wenn sich die Variationskoeffizienten des angenommenen Ankunftsprozesses und des tatsächlichen Ankunftsprozesses (aufgrund mißachteter Korrelationen) stark unterscheiden.

ρ_1	μ_2	UB	Sim	Δ (%)
0.1	1.0	1.0207	1.0265	0.5636
0.1	1.5	1.0053	1.0134	0.7967
0.1	2.0	0.9979	1.0081	1.0113
0.1	3.0	0.9969	1.0084	1.1353
0.2	1.0	1.0868	1.0896	0.2548
0.2	1.5	1.0221	1.0298	0.7476
0.2	2.0	0.9997	1.0073	0.7552
0.2	3.0	0.9871	0.9932	0.6140
0.3	1.0	1.1975	1.2001	0.2156
0.3	1.5	1.0515	1.0551	0.3401
0.3	2.0	1.0018	1.0078	0.5904
0.3	3.0	0.9698	0.9765	0.6881
0.4	1.0	1.3554	1.3559	0.0381
0.4	1.5	1.0899	1.0885	0.1338
0.4	2.0	1.0029	1.0062	0.3308
0.4	3.0	0.9482	0.9548	0.6862
0.5	1.0	1.5603	1.5484	0.7623
0.5	1.5	1.1370	1.1382	0.1098
0.5	2.0	1.0009	1.0029	0.2018
0.5	3.0	0.9152	0.9220	0.7417
0.6	1.0	1.8144	1.8012	0.7345
0.6	1.5	1.1849	1.1848	0.0149
0.6	2.0	0.9915	0.9961	0.4625
0.6	3.0	0.8750	0.8796	0.5222
0.7	1.0	2.1185	2.1053	0.6269
0.7	1.5	1.2249	1.2198	0.4161
0.7	2.0	0.9670	0.9677	0.0713
0.7	3.0	0.8185	0.8218	0.3983
0.8	1.0	2.4753	2.4596	0.6390
0.8	1.5	1.2255	1.2207	0.3921
0.8	2.0	0.9079	0.9088	0.1039
0.8	3.0	0.7435	0.7449	0.1933
0.9	1.0	2.8897	2.8794	0.3560
0.9	1.5	1.0927	1.1139	1.8979
0.9	2.0	0.7810	0.7926	1.4641
0.9	3.0	0.6421	0.6461	0.6133

Tabelle 5.10: Qualität der Abgangsprozesse

Die Überlegungen hinsichtlich der Erwartungshaltung an die Analyseresultate des Dekompositionsverfahrens für den betrachteten Typus von Fork/Join–Warteschlangennetzen werden im weiteren anhand verschiedener Experimente bestätigt. Dazu wird zunächst die Vermutung recht akkurater Resultate in azyklischen Netzen belegt. Anschließend zeigt die Betrachtung eines zyklischen Modells die in diesem Fall angesprochene Problematik der korrelierten Ankunftsprozesse auf.

5.3.1 Azyklische Fork/Join–Warteschlangennetze

In diesem Abschnitt werden einige Experimente an einem zyklusfreien Fork/Join–Warteschlangennetz durchgeführt. Die Experimente werden darlegen, daß das um das Analyseverfahren für das Upper–Bound Modell angereicherte Dekompositionsverfahren in diesem Kontext recht gute Resultate hinsichtlich der Momente der Gesamtdurchlaufzeit berechnet. Die Abbildung 5.8 skizziert das dazu herangezogene Modell. Das Modell besteht aus einem Fork/Join–Netz mit den parallelen Stationen S_1 und S_2 sowie aus den beiden PH/PH/1–∞ Stationen S_3 und S_4. Die Bedienzeitverteilungen der Kunden an den Stationen $S_1, \ldots, S_4$ sind je Station unabhängig identisch verteilt und durch ihre Raten $\mu_1, \ldots, \mu_4$ und quadratischen Variationskoeffizienten $c_1^2, \ldots, c_4^2$ charakterisiert. Die ebenfalls unabhängig identisch verteilten Zwischenankunftszeiten der Kunden an dem Fork/Join–Netz besitzen die Rate λ und den quadratischen Variationskoeffizienten c_A^2.

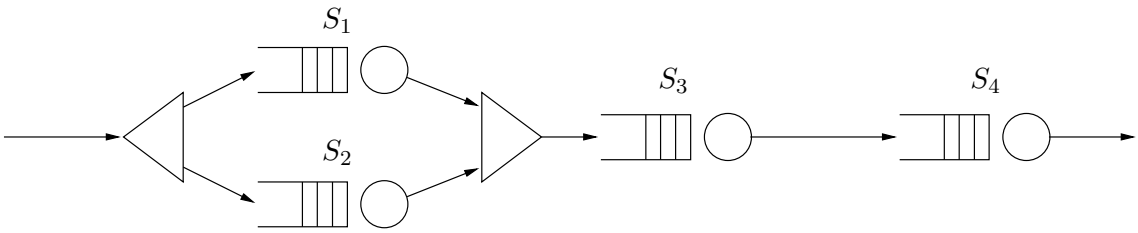

Abbildung 5.8: Azyklisches Fork/Join–Warteschlangennetz

Wie bereits zuvor erwähnt, sind aufgrund der geringen Approximationsfehler der Analyse der isolierten Fork/Join–Station ebenfalls qualitativ gute Analyseresultate des azyklischen Fork/Join–Warteschlangennetzes zu erwarten. Diese Vermutung wird im folgenden anhand dreier ausgewählter Experimentserien an obigem Modell belegt. Die Parameterwahl der Serien ist in der Tabelle 5.11 zusammengefaßt.

	(λ, c_A^2)	(μ_1, c_1^2)	(μ_2, c_2^2)	(μ_3, c_3^2)	(μ_4, c_4^2)
Experimentserie 1	(1.0;1.0)	$\left(\frac{\lambda_1}{\rho_1}; 0.5\right)$	(1.5μ_1;4.0)	(10.0;1.0)	(2.0;1.0)
Experimentserie 2	(1.0;1.0)	$\left(\frac{\lambda_1}{\rho_1}; 0.5\right)$	(1.5μ_1;4.0)	(2.0;1.0)	(2.0;1.0)
Experimentserie 3	(1.0;1.0)	$\left(\frac{\lambda_1}{\rho_1}; 0.5\right)$	(1.5μ_1;4.0)	($\frac{10.0}{9.0}$;1.0)	(2.0;1.0)

Tabelle 5.11: Parameter der Experimentserien

Die Werte von ρ_1 variieren jeweils zwischen 0.1 und 0.9, d.h. die Auslastung der Station S_1 nimmt dementsprechend Werte zwischen 0.1 und 0.9 an. Zur Beurteilung ihrer Qualität

werden die Analyseresultate des Dekompositionsverfahrens mit denen einer Simulation verglichen. Um möglichst vertrauenswürdige Ergebnisse zu erhalten, wurden extrem lange Simulationsläufe durchgeführt, so daß sich für die im folgenden dargestellten Erwartungswerte 95% Konfidenzintervalle der Breite 1% ergeben.

Die Analyseresultate der Experimentserien sind in den Tabellen 5.12, 5.13 und 5.14 dargestellt. Angegeben sind jeweils die Erwartungswerte der Durchlaufzeitverteilungen der isolierten Fork/Join–Station, der PH/PH/1/∞ Stationen S_3 und S_4 als auch des gesamten Netzes (QN). Die Spalten KW, Sim und Δ repräsentieren die Ergebnisse des Dekompositionsverfahrens, der Simulation bzw. die relative prozentuale Abweichung zwischen beiden Werten. Ebenso sind die Tabellen 5.15, 5.16 und 5.17 zu interpretieren, die die zugehörigen quadratischen Variationskoeffizienten der Durchlaufzeitverteilungen angeben.

	Fork/Join			S_3			S_4			QN		
ρ_1	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
0.1	0.15	0.15	0.33	0.11	0.11	0.20	1.00	1.00	0.26	1.26	1.26	0.18
0.2	0.33	0.33	0.48	0.11	0.11	0.17	1.01	1.02	1.13	1.45	1.46	-0.69
0.3	0.57	0.56	0.21	0.11	0.11	0.12	1.01	1.03	-2.03	1.69	1.71	-1.15
0.4	0.87	0.86	0.40	0.11	0.11	0.45	1.02	1.05	-2.91	2.00	2.03	-1.32
0.5	1.27	1.26	0.52	0.11	0.11	0.90	1.03	1.07	-3.50	2.41	2.44	-1.23
0.6	1.83	1.81	0.91	0.11	0.11	1.42	1.04	1.08	-3.43	2.98	3.00	-0.64
0.7	2.66	2.63	0.93	0.11	0.11	2.13	1.05	1.09	-2.85	3.82	3.83	-0.10
0.8	4.12	4.04	1.98	0.11	0.11	2.89	1.05	1.07	-1.05	5.28	5.21	1.38
0.9	8.12	7.81	3.97	0.11	0.11	3.86	1.02	1.00	2.07	9.26	8.92	3.75

Tabelle 5.12: Erwartungswerte der Durchlaufzeitverteilungen der Experimentserie 1

Insgesamt ist die Qualität der Resultate des Dekompositionsverfahrens sehr positiv zu beurteilen. Auffallend ist die Beobachtung, daß mit zunehmender Auslastung der Stationen S_1 und S_2 der prozentuale Fehler Δ der mittleren Durchlaufzeit aller Stationen sowie des gesamten Netzes anwächst. Die Ursache hierfür liegt, wie bereits angedeutet, in der fälschlichen Approximation des Abgangsprozesses der Fork/Join–Station durch einen stationären Erneuerungsprozeß. Dieser Fehler macht sich besonders im Fall hoher Auslastungen der Stationen S_1 und S_2 bemerkbar.

Ferner belegen die Experimentserien die Behauptung des vorangegangenen Abschnitts, daß eine besonders hoch ausgelastete PH/PH/1/∞ Station die fälschliche Approximation des Abgangsprozesses einer Fork/Join–Station durch einen stationären Erneuerungsprozeß filtert. Diese Tatsache wird anhand des Fehlers Δ der erwarteten Durchlaufzeit an der Station S_4 deutlich. Der Fehler wird offensichtlich mit steigender Auslastung $\rho_3 = 0.1, 0.5$ bzw. $\rho_3 = 0.9$ der Station S_3 in den Experimentserien 1, 2 bzw. 3 kleiner. Bestärkt wird diese Aussage zudem durch die Beobachtung des Fehlers Δ in den quadratischen Variationskoeffizienten der Durchlaufzeitverteilung an der Station S_4.

Der Vergleich der Spalte S_4 in der Tabelle 5.13 mit der Spalte S_3 der Tabelle 5.14 zeigt ferner, daß die mit $\rho_3 = 0.1$ sehr niedrig ausgelastete Station S_3 in der Experimentserie 1 den angesprochenen Fehler aufgrund der Approximation des Abgangsprozesses der Fork/Join Station ungefiltert an die nachfolgende Station durchreicht. Die Station S_4 in der Serie 1 und

die Station S_3 in der Serie 2 besitzen identische Auslastungen, und die Resultate sind ebenfalls auffallend nahe beieinander. Wiederum wird diese Beobachtung durch die Betrachtung der quadratischen Variationskoeffizienten bekräftigt.

Zuletzt sei die Frage betrachtet, ob sich die Eigenschaft des Upper–Bound Modells, den Erwartungswert der Durchlaufzeitverteilung der primären Fork/Join–Station nach oben abzuschätzen, auf die Durchlaufzeit des gesamten Fork/Join–Warteschlangennetzes übertragen läßt. Die Beobachtung der dritten Experimentserie scheint diese Frage zunächst sowohl hinsichtlich des Erwartungswertes und zudem sogar hinsichtlich der Variationskoeffizienten zu bejahen. Die Experimentserien 1 und 2 widerlegen dies jedoch. Anhand der Tabelle 5.14 wird deutlich, daß im Falle einer mittleren Auslastung ρ_1 der Station S_1 die erwartete Durchlaufzeit der Station S_3 (die ebenfalls mit $\rho_3 = 0.5$ eine mittlere Auslastung besitzt) tendenziell unterschätzt wird. Ebenso wird dieser Erwartungswert im Falle einer hohen Auslastung ρ_1 tendenziell überschätzt. Die gleiche Beobachtung ist aufgrund obiger Ausführungen wiederum für die Station S_4 der Experimentserie 1 gültig. In genau den beschriebenen Fällen wird auch die erwartete Gesamtdurchlaufzeit unter– bzw. überschätzt. Interessant ist, daß sich dieselbe Aussage auch anhand der Variationskoeffizienten ablesen ließe, was im vorhinein sicherlich nicht zu vermuten war.

Zusammenfassend sind die Resultate hinsichtlich der Erwartungswerte der Durchlaufzeitverteilungen der isolierten Stationen als auch des gesamten Fork/Join–Warteschlangennetzes sehr erfreulich. In den betrachteten Fällen liegen ferner die Approximationsfehler hinsichtlich der zugehörigen Variationskoeffizienten in einem akzeptablen Rahmen. Diese Beobachtung läßt sich jedoch nicht verallgemeinern, da die Vernachlässigung von Abhängigkeiten unter den Durchlaufzeitverteilungen der isolierten Stationen zur Ermittlung der Durchlaufzeitverteilung des gesamten Netzes zu deutlichen Fehlern des zweiten Moments führt.

	Fork/Join			S_3			S_4			QN		
ρ_1	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
0.1	0.15	0.15	0.25	1.00	1.00	0.06	1.00	1.00	0.07	2.15	2.15	0.02
0.2	0.33	0.33	0.54	1.01	1.02	-1.05	1.00	1.01	-0.33	2.34	2.35	-0.52
0.3	0.57	0.56	0.13	1.01	1.03	-1.80	1.01	1.01	-0.47	2.59	2.61	-0.87
0.4	0.87	0.86	0.40	1.02	1.05	-2.69	1.02	1.03	-0.86	2.91	2.94	-1.14
0.5	1.27	1.26	0.33	1.03	1.07	-3.40	1.03	1.04	-1.30	3.33	3.37	-1.36
0.6	1.83	1.81	0.89	1.05	1.08	-3.31	1.03	1.05	-1.55	3.90	3.94	-0.91
0.7	2.66	2.63	1.08	1.05	1.08	-2.64	1.04	1.05	-1.04	4.75	4.76	-0.23
0.8	4.12	4.05	1.58	1.05	1.06	-0.70	1.04	1.04	0.09	6.21	6.16	0.90
0.9	8.12	7.78	4.38	1.02	1.00	2.21	1.02	1.00	1.61	10.16	9.78	3.88

Tabelle 5.13: Erwartungswerte der Durchlaufzeitverteilungen der Experimentserie 2

	Fork/Join			S_3			S_4			QN		
ρ_1	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
0.1	0.15	0.15	0.29	9.02	9.01	0.18	1.00	1.00	0.02	10.17	10.15	0.17
0.2	0.33	0.33	0.55	9.09	8.91	1.99	1.00	1.00	0.05	10.42	10.24	1.76
0.3	0.57	0.56	0.21	9.21	9.12	0.95	1.00	1.00	0.14	10.78	10.69	0.83
0.4	0.87	0.87	0.32	9.36	9.21	1.68	1.00	1.00	0.03	11.24	11.08	1.42
0.5	1.27	1.26	0.34	9.55	9.39	1.73	1.01	1.01	0.05	11.83	11.66	1.43
0.6	1.83	1.81	0.67	9.75	9.48	2.82	1.01	1.01	0.03	12.58	12.30	2.28
0.7	2.66	2.63	0.95	9.91	9.61	3.17	1.01	1.01	0.13	13.58	13.25	2.48
0.8	4.12	4.05	1.63	9.91	9.65	2.74	1.01	1.01	0.02	15.04	14.71	2.24
0.9	8.12	7.87	3.17	9.38	9.07	3.37	1.00	1.00	0.06	18.50	17.95	3.09

Tabelle 5.14: Erwartungswerte der Durchlaufzeitverteilungen der Experimentserie 3

	Fork/Join			S_3			S_4			QN		
ρ_1	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
0.1	1.10	1.09	0.83	1.00	1.00	-0.47	1.00	1.00	0.19	0.66	0.65	0.14
0.2	1.13	1.12	0.98	1.00	1.01	-1.30	1.00	1.00	-0.36	0.55	0.55	1.36
0.3	1.12	1.12	0.55	1.00	1.02	-1.88	1.00	1.02	-1.51	0.49	0.51	-4.02
0.4	1.09	1.09	0.32	1.00	1.02	-2.22	1.00	1.04	-3.97	0.47	0.51	-7.09
0.5	1.04	1.03	0.29	1.00	1.02	-2.09	1.00	1.07	-6.54	0.47	0.52	-8.54
0.6	0.96	0.96	0.76	1.00	1.02	-2.17	1.00	1.10	-9.00	0.49	0.53	-8.18
0.7	0.87	0.86	1.15	1.00	1.02	-1.91	1.00	1.14	-12.03	0.50	0.53	-6.60
0.8	0.78	0.76	2.28	1.00	1.02	-1.60	1.00	1.15	-13.13	0.51	0.52	-2.13
0.9	0.76	0.74	1.78	1.00	1.01	-1.24	1.00	1.16	-14.05	0.59	0.58	1.86

Tabelle 5.15: Quadratische Variationskoeffizienten der Durchlaufzeitverteilungen in der Experimentserie 1

	Fork/Join			S_3			S_4			QN		
ρ_1	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
0.1	1.10	1.09	1.12	1.00	1.00	-0.14	1.00	1.00	0.01	0.44	0.44	0.03
0.2	1.13	1.12	1.38	1.00	1.01	-0.52	1.00	1.00	0.10	0.39	0.39	0.79
0.3	1.12	1.12	0.39	1.00	1.02	-1.72	1.00	1.00	-0.24	0.36	0.37	-2.44
0.4	1.09	1.09	0.17	1.00	1.04	-3.85	1.00	1.00	-0.48	0.34	0.36	-5.45
0.5	1.04	1.04	0.11	1.00	1.07	-6.37	1.00	1.01	-1.43	0.34	0.37	-8.25
0.6	0.96	0.96	0.60	1.00	1.11	-9.70	1.00	1.03	-2.60	0.35	0.39	-9.66
0.7	0.87	0.86	1.24	1.00	1.14	-12.28	1.00	1.04	-3.44	0.37	0.40	-8.50
0.8	0.78	0.77	1.17	1.00	1.16	-14.04	1.00	1.05	-4.31	0.40	0.42	-5.08
0.9	0.76	0.74	2.06	1.00	1.16	-14.05	1.00	1.04	-3.92	0.50	0.49	2.42

Tabelle 5.16: Quadratische Variationskoeffizienten der Durchlaufzeitverteilungen in der Experimentserie 2

	Fork/Join			S_3			S_4			QN		
ρ_1	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
0.1	1.10	1.09	0.68	1.00	0.97	2.70	1.00	1.00	0.29	0.80	0.78	2.73
0.2	1.13	1.12	1.13	1.00	0.98	2.26	1.00	1.00	0.02	0.77	0.75	2.77
0.3	1.12	1.12	0.38	1.00	1.00	0.13	1.00	1.00	0.02	0.74	0.74	0.56
0.4	1.09	1.09	0.41	1.00	0.98	2.33	1.00	1.00	0.03	0.71	0.69	3.18
0.5	1.04	1.04	0.07	1.00	0.96	3.96	1.00	1.00	0.16	0.67	0.64	5.06
0.6	0.96	0.96	0.54	1.00	0.94	6.16	1.00	1.00	0.01	0.63	0.58	7.76
0.7	0.87	0.86	1.21	1.00	0.94	6.95	1.00	1.00	0.00	0.57	0.53	8.19
0.8	0.78	0.77	1.62	1.00	0.92	8.24	1.00	1.00	0.09	0.50	0.46	7.40
0.9	0.76	0.76	0.68	1.00	0.94	6.33	1.00	1.00	0.50	0.41	0.40	0.57

Tabelle 5.17: Quadratische Variationskoeffizienten der Durchlaufzeitverteilungen in der Experimentserie 3

5.3.2 Zyklische Fork/Join–Warteschlangennetze

Die Struktur des in diesem Abschnitt betrachteten zyklischen Fork/Join–Warteschlangennetzes ist in der Abbildung 5.9 dargestellt. Das Modell entspricht in wesentlichen Teilen dem zuvor behandelten azyklischen Modell. Im Unterschied zu diesem erfolgt jedoch nach der Bedienung eines Kunden an der Station S_3 mit der Wahrscheinlichkeit p eine Rückkehr zu der aus S_1 und S_2 aufgebauten Fork/Join–Station. Die Anzahl der Durchläufe durch den aus der Fork/Join–Station und der Station S_3 gebildeten Zyklus ist geometrisch mit dem Parameter p verteilt. Folglich ist $\frac{1}{1-p}$ die mittlere Anzahl an Durchläufen, die ein Kunde in dem Zyklus verbringt.

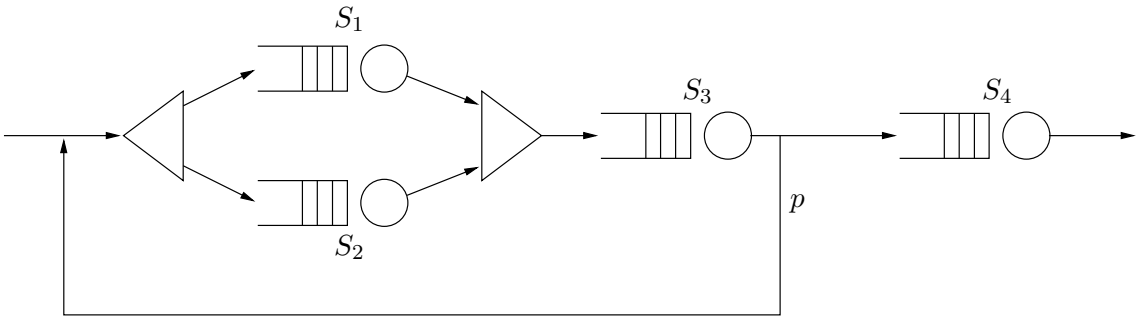

Abbildung 5.9: Zyklisches Fork/Join-Warteschlangennetz

Die Experimente, die im folgenden an diesem Modell beschrieben werden, sind in der Tabelle 5.18 zusammengefaßt. Das Paar (λ, c_A^2) gibt die Rate und den quadratischen Variationskoeffizienten des phasenverteilten Ankunftsprozesses wieder, und dementsprechend spezifizieren die Paare $(\mu_i, c_i^2), i = 1, \ldots, 4$ die Raten und quadratischen Variationskoeffizienten der phasenverteilten Bedienprozesse an den Stationen $S_1, \ldots, S_4$. In allen Experimentserien wird der Parameter p derart variiert, daß der Zyklus im Mittel zwischen einem und neun mal durchlaufen wird. Die Wahl der Rate μ_3 hat zur Folge, daß die Station S_3 in den Experimenten der Serien 1,2 bzw. 3 die konstanten Auslastungen 0.1, 0.5 bzw. 0.9 hat.

	(λ, c_A^2)	(μ_1, c_1^2)	(μ_2, c_2^2)	(μ_3, c_3^2)	(μ_4, c_4^2)
Experimentserie 1	(1.0;1.0)	(10.0;0.5)	(20.0;4.0)	$\left(\frac{10.0}{(1-p)}; 1.0\right)$	(2.0;1.0)
Experimentserie 2	(1.0;1.0)	(10.0;0.5)	(20.0;4.0)	$\left(\frac{2.0}{(1-p)}; 1.0\right)$	(2.0;1.0)
Experimentserie 3	(1.0;1.0)	(10.0;0.5)	(20.0;4.0)	$\left(\frac{10.0}{9.0(1-p)}; 1.0\right)$	(2.0;1.0)

Tabelle 5.18: Konfiguration der Experimente

Die Tabelle 5.19 stellt die Resultate der Experimentserie 1 dar. Gezeigt ist jeweils der Erwartungswert der Durchlaufzeit der isolierten Fork/Join–Station, der Stationen S_3 und S_4 sowie des gesamten Netzes (QN). Die Spalten KW, Sim bzw. Δ geben die anhand des Dekompositionsverfahrens und anhand einer Simulation ermittelten Werte an sowie die relative prozentuale Abweichung beider. Die erste Spalte gibt die mittlere Anzahl an Zyklusdurchläufen an. Wie bereits in der Einführung zu diesem Abschnitt erwartet, resultieren insbesondere im Fall einer hohen Anzahl an Zyklusdurchläufen erhebliche Approximationsfehler hinsichtlich des Erwartungswertes der Durchlaufzeit der Fork/Join–Station und somit auch des gesamten

	Fork/Join			S_3			S_4			QN		
	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
1	0.13	0.13	0.66	0.11	0.11	0.51	1.00	0.99	1.10	1.24	1.23	1.00
2	0.15	0.15	0.66	0.06	0.05	1.26	1.00	0.99	0.85	1.40	1.40	0.65
3	0.16	0.17	0.84	0.04	0.04	2.58	1.00	1.01	0.85	1.60	1.61	0.75
4	0.19	0.19	2.96	0.03	0.03	3.20	1.00	1.00	0.26	1.85	1.87	0.94
5	0.21	0.22	4.20	0.02	0.02	3.97	1.00	0.97	2.58	2.18	2.20	0.76
6	0.25	0.28	9.84	0.02	0.02	3.90	1.00	0.99	1.43	2.63	2.78	5.45
7	0.31	0.36	13.36	0.02	0.02	3.31	1.00	0.99	0.68	3.31	3.64	9.01
8	0.43	0.54	20.18	0.01	0.01	2.11	0.99	0.99	0.81	4.52	5.37	15.85
9	0.74	1.01	26.83	0.01	0.01	0.83	0.99	0.98	0.80	7.77	10.20	23.87

Tabelle 5.19: Resultate der Experimentserie 1

Netzes. Die Ursache hierfür ist bei der Betrachtung des quadratischen Variationskoeffizienten der Zwischenabgangszeiten der Station S_3 schnell gefunden. Die Tabelle 5.20 stellt den anhand des Dekompositionsverfahrens errechneten und den in der Simulation beobachteten quadratischen Variationskoeffizienten dar. Die Spalten sind wie gewohnt zu interpretieren. Die Mißachtung von Korrelationen, die Aufgrund des Zyklus in den Ankunftsprozessen der

Loops	KW	Sim	Δ (%)
1	1.00	1.01	0.50
2	1.00	2.40	58.28
3	1.00	3.38	70.32
4	1.00	3.97	74.69
5	1.00	4.19	76.11
6	0.99	4.01	75.30
7	0.96	3.66	73.65
8	0.91	2.85	68.13
9	0.79	1.84	56.98

Tabelle 5.20: Quadrat. Variationskoeffizient der Zwischenabgangszeiten der Station S_3

Fork/Join–Station und der Station S_3 auftreten, führt zu erheblichen Fehlern in den Variationskoeffizienten der Schnittstellenprozesse und damit auch in den Erwartungswerten der Durchlaufzeitverteilungen. Ferner ist die Anzahl der Zyklusdurchläufe ein Maß für die Auslastungen der in den Zyklus integrierten Stationen. Die Tabelle 5.20 macht somit deutlich, daß Stationen mit mittlerer Auslastung stärkere Korrelationen zur Folge haben als niedrig und hoch ausgelastete Stationen. Andererseits steigt jedoch der Einfluß des Variationskoeffizienten des Ankunftsprozesse auf den Erwartungswert der Durchlaufzeitverteilung mit der Auslastung einer PH/PH/1-∞ Station an. Daher sind die Auswirkungen des Approximationsfehlers auf die mit 0.1 sehr niedrig ausgelastete Station S_3 sehr gering (vgl. Tabelle 5.19).

Die erste Experimentserie bestätigt ferner die geäußerte Vermutung, daß in zyklischen Netzen die Eigenschaft des Upper–Bound Modells, die mittlere Durchlaufzeit des primären Modells nach oben hin abzuschätzen, keineswegs zwingend erhalten bleibt. Folglich bildet auch die anhand des Dekompositionsverfahrens errechnete mittlere Gesamtdurchlaufzeit keine obere

Schranke für den tatsächlichen Wert.

In einer zweiten Experimentserie wurde die Auslastung der Station S_3 auf 0.5 erhöht. Die Resultate dieser Serie sind in der Tabelle 5.21 gesammelt. Der Vergleich der Spalte S_3 der

	Fork/Join			S_3			S_4			QN		
	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
1	0.13	0.13	0.70	1.00	0.98	1.62	1.00	0.99	1.31	2.13	2.10	1.42
2	0.15	0.15	0.86	0.50	0.51	1.45	1.00	0.99	0.95	2.29	2.30	0.44
3	0.16	0.16	0.57	0.33	0.34	1.46	1.00	1.00	0.26	2.49	2.51	0.67
4	0.18	0.19	1.19	0.25	0.25	0.62	1.00	1.01	0.54	2.74	2.75	0.52
5	0.21	0.22	2.72	0.20	0.19	3.16	1.00	1.00	0.41	3.07	3.07	0.14
6	0.25	0.27	7.15	0.17	0.16	3.46	1.00	0.99	1.24	3.52	3.60	2.31
7	0.32	0.36	10.91	0.14	0.13	6.06	1.00	0.99	0.70	4.20	4.41	4.74
8	0.44	0.53	17.20	0.12	0.11	6.70	1.00	0.99	0.67	5.46	6.12	10.90
9	0.78	1.01	22.49	0.10	0.10	5.16	0.99	0.99	0.53	8.92	10.91	18.18

Tabelle 5.21: Resultate der Experimentserie 2

Tabellen 5.19 und 5.21 zeigt, daß sich der Approximationsfehler aufgrund der höheren Auslastung der Station S_3 in der zweiten Serie deutlich stärker auf den Erwartungswert der Durchlaufzeit der isolierten Station auswirkt. Gleichzeitig sinkt jedoch der Fehler hinsichtlich der Fork/Join–Station und des gesamten Netzes wegen der insgesamt höheren Auslastung der Stationen innerhalb des Zyklus.

Noch deutlicher ist der Effekt geringer Korrelation im Falle hoher Auslastungen in der dritten Experimentserie zu erkennen. Die Auslastung der Station S_3 ist mit 0.9 sehr hoch gewählt. Die Tabelle 5.22 verdeutlicht, daß in diesem Fall die Approximationsfehler in der mittleren Durchlaufzeit sowohl bzgl. der isolierten Stationen als auch bzgl. des gesamten Netzes im Vergleich zu den ersten beiden Experimentserien deutlich geringer ausfallen. Zu erklären

	Fork/Join			S_3			S_4			QN		
	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)	KW	Sim	Δ (%)
1	0.13	0.13	0.14	9.00	9.18	1.92	1.00	1.00	0.13	10.13	10.31	1.72
2	0.15	0.15	0.86	4.50	4.52	0.55	1.00	1.00	0.12	10.29	10.35	0.51
3	0.16	0.16	0.70	3.00	2.99	0.28	1.00	1.00	0.23	10.50	10.48	0.21
4	0.19	0.18	0.82	2.25	2.27	0.80	1.00	1.00	0.06	10.75	10.81	0.53
5	0.21	0.21	0.80	1.80	1.81	0.75	1.00	1.00	0.16	11.07	11.13	0.54
6	0.25	0.25	0.33	1.49	1.51	0.99	1.00	1.00	0.01	11.49	11.57	0.68
7	0.33	0.32	1.96	1.26	1.25	1.01	1.00	1.00	0.16	12.14	12.01	1.09
8	0.45	0.47	4.02	1.07	1.12	3.91	1.00	1.00	0.07	13.15	13.67	3.76
9	0.83	0.90	7.29	0.89	0.89	0.59	1.00	0.99	0.38	16.52	17.05	3.07

Tabelle 5.22: Resultate der Experimentserie 3

ist diese Beobachtung wiederum anhand des quadratischen Variationskoeffizienten der Zwischenabgangszeiten der Station S_3. Der Vergleich der folgenden Tabelle 5.23 mit der Tabelle 5.20 zeigt erheblich geringere Korrelationen in den Schnittstellenprozessen der dritten Expe-

rimentserie. Diese Tatsache ist auf die sehr hohe Auslastung der Station S_3 zurückzuführen.

Loops	KW	Sim	Δ (%)
1	1.00	1.00	0.09
2	1.00	1.15	13.40
3	1.00	1.26	20.80
4	1.00	1.34	25.52
5	1.00	1.37	27.05
6	1.00	1.38	27.50
7	1.00	1.35	26.09
8	0.99	1.27	21.84
9	0.98	1.17	16.85

Tabelle 5.23: Quadrat. Variationskoeffizient der Zwischenabgangszeiten der Station S_3

Zusammenfassend bleibt festzuhalten, daß in zyklischen Netzen mit deutlichen Approximationsfehlern des Dekompositionsverfahrens zu rechnen ist. Diese resultieren aus der Mißachtung von Korrelationen in den Ankunftsprozessen. Auswirkungen auf die Korrelation haben zum einen die Auslastungen der Stationen innerhalb der Zyklen und zum anderen die Länge der Zyklen. Ohne daß dies detailliert dargestellt wurde, ist in langen Zyklen, d.h. im Falle vieler Stationen innerhalb eines Zyklus, mit geringeren Korrelationen zu rechnen als in kurzen Zyklen. Ferner bleibt festzuhalten, daß in derartigen Fork/Join–Warteschlangennetzen die Eigenschaft des Upper–Bound Modells, die erwartete Durchlaufzeit des primären Fork/Join Modells nach oben abzuschätzen, im allgemeinen nicht erhalten bleibt.

Kapitel 6

Erweiterung auf allgemeine Fork/Join–Netze

Das in Kapitel 4.1 vorgestellte Lösungsverfahren erlaubt die Analyse einer recht eingeschränkten Klasse von Fork/Join–Netzen nämlich solcher, die eine feste Anzahl paralleler PH/PH/1/∞ Systeme mit FCFS–Bedienung synchronisieren. Im folgenden wird dieser Stationstyp mit dem Begriff *einfache Fork/Join–Station* bezeichnet. In der Praxis treten jedoch häufig Fälle auf, die die Synchronisation deutlich komplexerer Netzstrukturen erfordern. Eine derartige allgemeine Fork/Join–Station ist in der Abbildung 6.1 skizziert.

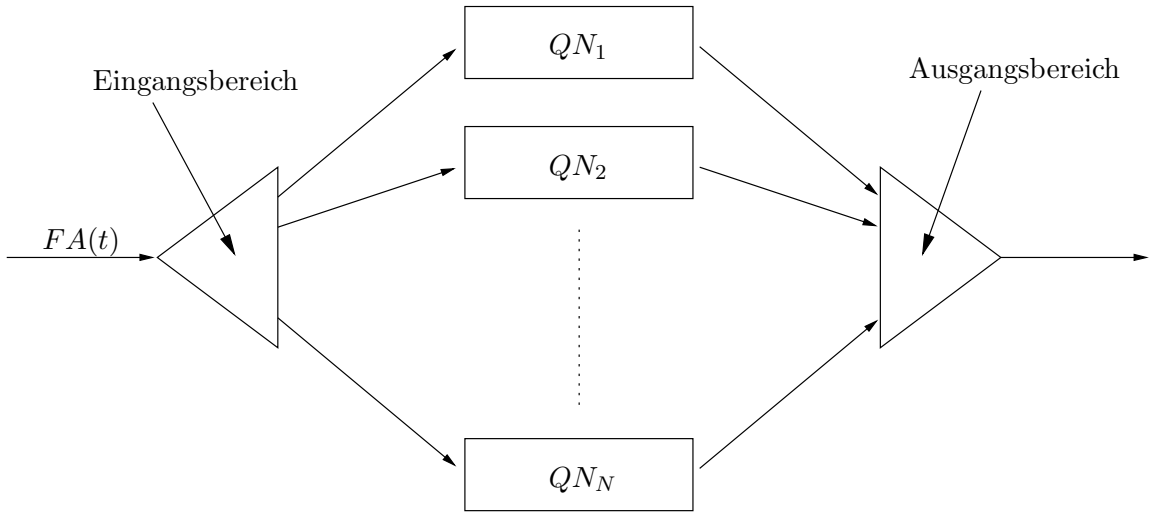

Abbildung 6.1: Allgemeines Fork/Join–Netz

Der Aufbau des Modells ist dem der einfachen Fork/Join–Station sehr ähnlich. Das Modell besteht aus einem Eingangs- und Ausgangsbereich sowie einer festen Anzahl N von parallelen Netzen. Im Unterschied zu dem einfachen Modell sind die $QN_i, i = 1, \ldots, N$ jedoch allgemeine Warteschlangennetze mit einem ausgezeichneten Eintritts- und Austrittsknoten. Auf die konkrete Ausprägung, die die QN_i besitzen dürfen, wird in Abschnitt 6.4 eingegangen.

Aus dynamischer Sicht wird das Modell durch Aufträge belastet, die gemäß der (phasenverteilten) Zwischenankunftszeiten $FA(t)$ im Eingangsbereich der Fork/Join–Station ankommen. Nach Ankunft eines Auftrags wird dieser unmittelbar in genau N Teilaufträge zerlegt. Jedes der N Warteschlangennetze erhält genau einen Teilauftrag. Diese durchlaufen die QN_i und warten anschließend im Ausgangsbereich. Ein Auftrag gilt als erfüllt, wenn all seine N Teilaufträge vollständig bearbeitet sind und somit im Ausgangsbereich warten.

Eine direkte Behandlung der allgemeinen Fork/Join–Station mit dem in Kapitel 4.1 vorgestellten Analyseansatz für einfache Fork/Join–Stationen würde zu einer Explosion des Zustandsraumes des zugrundeliegenden QBDs führen und ist somit unter Aufwandsaspekten wie Rechenzeit und Speicherplatzbedarf nicht realisierbar. In dieser Arbeit wird daher ein Analyseansatz verfolgt, der derart große Zustandsräume vermeidet. Statt dessen wird eine erweiterte Fork/Join–Station durch eine einfache Fork/Join–Station ersetzt, die das ursprüngliche Netz in seinem Verhalten approximiert. Diese einfache Fork/Join–Station kann dann mit dem in Kapitel 4.1 vorgestellten Verfahren analysiert werden.

Die Idee der Approximation besteht darin, jedes einzelne der Warteschlangennetze QN_i (vgl. Abb. 6.1) durch spezielle PH/PH/1/∞–Systeme zu aggregieren. Die Aggregate haben die Eigenschaft, daß sie unter der Zwischenankunftszeitverteilung $FA(t)$ die Durchlaufzeitverteilungen der Warteschlangennetze QN_i approximieren. Da die Anpassung der gesamten Verteilung deutlich zu schwierig ist, werden wiederum ausschließlich die ersten beiden Momente betrachtet. Ferner wird davon ausgegangen, daß sich sowohl die Zwischenankunftszeiten als auch die Bedienzeiten in Abhängigkeit ihres Variationskoeffizienten durch die in Anhang A skizzierten speziellen Phasenverteilungen ($\mathcal{H}$–Verteilung) approximieren lassen. Je QN_i liegt die zu leistende Arbeit somit darin, unter Kenntnis der $\mathcal{H}$-verteilten Zwischenankunftszeiten der allgemeinen Fork/Join–Station und der Kenntnis der ersten beiden Momente der Durchlaufzeitverteilung des Warteschlangennetzes QN_i die $\mathcal{H}$–verteilten Bedienzeiten eines geeigneten $\mathcal{H}/\mathcal{H}/1\infty$–Aggregats zu bestimmen, das dieselben ersten beiden Momente der Durchlaufzeitverteilung besitzt. Diese Aggregierung wird je QN_i gelingen, wenn der jeweilige Erwartungswert der Durchlaufzeit in einem durch die Varianz der Durchlaufzeit (und durch die Parameter des Ankunftsprozesses) bestimmten Bereich liegt. Eine konkrete hinreichende Bedingung für die Existenz eines Aggregats wird im nachfolgenden Abschnitt formuliert.

Die Analyse einer erweiterten Fork/Join–Station kann somit folgendermaßen skizziert werden:

1. Bestimme für jedes Warteschlangennetz QN_i den Erwartungswert und die Varianz der Durchlaufzeitverteilung unter den $\mathcal{H}$–verteilten Zwischenankunftszeiten $FA(t)$.
2. Bestimme für jedes QN_i ein geeignetes $\mathcal{H}/\mathcal{H}/1/\infty$–System, das dieselben ersten beiden Momente der Durchlaufzeitverteilung besitzt.
3. Ersetze die Warteschlangennetze QN_i durch die errechneten Aggregate, und analysiere das Ersatznetz mit dem in Kapitel 4.1 vorgestellten Verfahren für einfache Fork/Join–Stationen.

In den folgenden Abschnitten wird die Berechnung der $\mathcal{H}/\mathcal{H}/1/\infty$–Aggregate erläutert. Dazu wird zunächst ein in gewissem Sinne allgemeiner Fall betrachtet, für den sich ein Iterationsschema zur Ermittlung der $\mathcal{H}$–verteilten Bedienzeiten angeben läßt. Anschließend werden

Spezialfälle betrachtet, in denen entweder der Variationskoeffizient der Zwischenankunftszeiten oder der Variationskoeffizient der Durchlaufzeiten oder beide 1 sind. In diesen Fällen läßt sich eine Lösung algebraisch ermitteln.

6.1 $\mathcal{H}/\mathcal{H}/1/\infty$–Aggregate

In diesem Abschnitt wird ein $\mathcal{H}/\mathcal{H}/1/\infty$–System mit FCFS–Bediendisziplin betrachtet. Dieses System zeichnet sich durch eine $\mathcal{H}$–verteilte Zwischenankunfts– und Bedienzeitverteilung und einen unbeschränkten Warteraum aus. Da $\mathcal{H}$–Verteilungen spezielle Phasenverteilungen sind, lassen sich anhand der Ausführungen zu Abschnitt 3.2 verschiedene stationäre Leistungsmaße dieses Stationstyps durch Analyse des zugrundeliegenden QBDs errechnen. Insbesondere sind die ersten beiden Momente der Durchlaufzeitverteilung bestimmbar. Sind die Zwischenankunftszeitverteilung und die Bedienzeitverteilung durch ihre Raten und Variationskoeffizienten (λ, c_A) und (μ, c_B) beschrieben, so ergeben sich der Erwartungswert und die Varianz der Durchlaufzeitverteilung als Funktion dieser Parameter. Konkret sei der Erwartungswert durch die Funktion $E_D(\lambda, c_A, \mu, c_B)$ und die Varianz durch die Funktion $Var_D(\lambda, c_A, \mu, c_B)$ beschrieben.

Das Ziel dieses Abschnittes besteht umgekehrt darin, aus der Kenntnis der $\mathcal{H}$–verteilten Zwischenankunftszeit und der Kenntnis der (im allgemeinen nicht $\mathcal{H}$–verteilten) Durchlaufzeit auf die $\mathcal{H}$–Verteilung der Bedienzeit zu schließen. Dazu sei nochmals angemerkt, daß sich die Momente der Durchlaufzeit aus der Analyse der separat betrachteten QN_i aus Abbildung 6.1 unter der gemeinsamen Zwischenankunftszeitverteilung und unter Anwendung des Dekompositionsverfahrens ergeben. Es wird gezeigt, daß eine derartige Bedienzeitverteilung existiert, wenn der Erwartungswert der Durchlaufzeit in einem durch die Varianz der Durchlaufzeit bestimmten Bereich liegt. Ferner wird ein iteratives Verfahren zur Bestimmung der Bedienzeit entwickelt.

Mit diesen Vorbereitungen läßt sich das Ziel dieses Abschnittes formalisieren. Die Aufgabe besteht darin, unter der Kenntnis der Parameter (λ, c_A) der Zwischenankunftszeitverteilung und des Erwartungswertes ED und der Varianz VD der Durchlaufzeitverteilung geeignete Parameter (μ, c_B) der $\mathcal{H}$–verteilten Bedienzeit eines $\mathcal{H}/\mathcal{H}/1/\infty$ Systems derart zu bestimmen, daß gilt:

$$E_D(\lambda, c_A, \mu, c_B) = ED \tag{6.1}$$

$$Var_D(\lambda, c_A, \mu, c_B) = VD \tag{6.2}$$

Die Idee des Iterationsverfahrens besteht darin, ausgehend von einem Startwert μ_0 einen Wert c_0 derart zu bestimmen, daß das Paar (μ_0, c_0) die Gleichung 6.2 erfüllt. Anschließend wird ein Wert μ_1 so bestimmt, daß das Paar (μ_1, c_0) die Gleichung 6.1 erfüllt. Im folgenden wird gezeigt, daß die wiederholte Durchführung dieser Schritte sicher zu einer Lösung der obigen Aufgabe führt, wenn die bereits angesprochene und in den nachfolgenden Ausführungen formulierte hinreichende Bedingung für die Existenz eines Aggregats erfüllt ist. Dazu werden zunächst unter der Annahme fester Werte λ und c_A die folgenden zweidimensionalen Funktionen F_1 und F_2 definiert.

Definition 6.1 *Seien die Rate λ und der Variationskoeffizient c_A der Zwischenankunftszeitverteilung eines $\mathcal{H}/\mathcal{H}/1/\infty$–Systems fest gewählt. Dann sind die Funktionen $F_1(x,y)$ und $F_2(x,y)$ folgendermaßen definiert:*

$$\begin{aligned} F_1(x,y) &:= E_D(\lambda, c_A, x, y) \\ F_2(x,y) &:= Var_D(\lambda, c_A, x, y) \end{aligned}$$

Zur Konstruktion des Iterationsschemas sind zunächst einige Monotonieaussagen über die Funktionen F_1 und F_2 zu beweisen.

Lemma 6.1 *Für beliebige, fest gewählte $c_B \in]0,\infty[$ sind die Funktionen*

1. $F_1 : \{(x, c_B) : x \in]\lambda, \infty[\} \rightarrow]\infty, 0[$ *und*
2. $F_2 : \{(x, c_B) : x \in]\lambda, \infty[\} \rightarrow]\infty, 0[$

stetig und streng monoton fallend. Im Fall $0 < c_B < 1$ wird zudem verlangt, daß sich die Anzahl der Phasen der Bedienzeitverteilung nicht ändert.

<u>Beweis:</u>

1. *Zum Beweis der strengen Monotonie der Funktionen $F_1(x, c_B)$ und $F_2(x, c_B)$ für feste $c_B \in]0,\infty[$ sei zunächst angemerkt, daß sich der Erwartungswert bzw. die Varianz der Durchlaufzeitverteilung eines $\mathcal{H}/\mathcal{H}/1/\infty$ Systems aus der Summe der Erwartungswerte bzw. der Varianzen der Bedienzeitverteilung und der Wartezeitverteilung ergeben. Ist die Wartezeitverteilung durch die Funktion $W(x, c_B)$ gegeben, so lassen sich die Funktionen F_1 und F_2 folgendermaßen darstellen:*

$$F_1(x, c_B) = \frac{1}{x} + E[W(x, c_B)] \tag{6.3}$$

$$F_2(x, c_B) = \frac{c_B^2}{x^2} + Var(W(x, c_B)) \tag{6.4}$$

Seien also Bedienraten $\lambda < \mu_1 < \mu_2 < \infty$ betrachtet. Dann folgt nach Satz A.9 für die durch (μ_1, c_B) und (μ_2, c_B) eindeutig charakterisierten $\mathcal{H}$–verteilten Bedienzeiten B_1 und B_2 die Beziehung $B_1 \overset{(1)}{\geq} B_2$. Mit Satz A.1 folgt unmittelbar für die Verteilungen der Wartezeiten $W(\mu_1, c_B) \overset{(2)}{\geq} W(\mu_2, c_B)$ und ferner mit Satz A.2 $E[W(\mu_1, c_B)] \geq E[W(\mu_2, c_B)]$ und $Var(W(\mu_1, c_B)) \geq Var(W(\mu_2, c_B))$. Zusammenfassend resultiert daraus die strenge Monotonie der Funktionen $F_1(x, c_B)$ und $F_2(x, c_B)$ folgendermaßen:

$$F_1(\mu_1, c_B) = \frac{1}{\mu_1} + E[W(\mu_1, c_B)] \geq \frac{1}{\mu_1} + E[W(\mu_2, c_B)] > \frac{1}{\mu_2} + E[W(\mu_2, c_B)]$$

$$F_2(\mu_1, c_B) = \frac{c_B^2}{\mu_1^2} + Var\left(W(\mu_1, c_B)\right) \geq \frac{c_B^2}{\mu_1^2} + Var\left(W(\mu_2, c_B)\right) > \frac{c_B^2}{\mu_2^2} + Var\left(W(\mu_2, c_B)\right)$$

2. *Unter der Voraussetzung, daß* $E[W(x, c_B)]$ *und* $Var(W(x, c_B))$ *für festes* $c_B \in]0, \infty[$ *auf dem Intervall* $x \in]0, \infty[$ *stetig sind, sind* $F_1(x, c_B)$ *und* $F_2(x, c_B)$ *als Summe stetiger Funktionen (vgl. 6.3) ebenfalls auf dem Intervall* $x \in]0, \infty[$ *stetig. Auf die Stetigkeit von* $E[W(x, c_B)]$ *und* $Var(W(x, c_B))$ *wird an dieser Stelle nicht weiter eingegangen.*

3. *Die Bildbereiche der Funktionen* $F_1(x, y)$ *und* $F_2(x, y)$ *für feste* $y = c_B \in]0, \infty[$ *resultieren aus den folgenden Überlegungen:*

   Aus Systemsicht bedeuten sehr hohe Bedienraten eine extrem niedrige Auslastung. Der Bediener ist folglich kaum beschäftigt und bedient ankommende Kunden unmittelbar. Die ersten beiden Momente der Durchlaufzeit entsprechen folglich nahezu den ersten beiden Momenten der Bedienzeit. Somit gelten die folgenden Gleichungen:

$$\lim_{x \to \infty} F_1(x, c_B) = \lim_{x \to \infty} \frac{1}{x} = 0 \tag{6.5}$$

$$\lim_{x \to \infty} F_2(x, c_B) = \lim_{x \to \infty} \frac{c_B^2}{x^2} = 0 \tag{6.6}$$

   Andererseits ist die Auslastung des Systems im Falle $\mu \to \lambda$ *extrem hoch. Der Bediener ist folglich (nahezu) ständig beschäftigt. Demzufolge konvergieren die ersten beiden Momente der Zwischenabgangszeiten gegen die ersten beiden Momente der Bedienzeitverteilung. Ist die* $\mathcal{H}$*–verteilte Bedienzeit durch die Phasendarstellung* (B, β) *repräsentiert, so folgt im Fall* $\mu \to \lambda$ *für den Erwartungswert* $E[X]$ *der Zwischenabgangszeiten:*

$$E[X] = -\beta B^{-1} e$$

   (vgl. hierzu die Darstellung der Momente einer Phasenverteilung in Anhang A, Gleichungen (A.2) und (A.3)). Verglichen mit der absorbierenden Markovkette zur Ermittlung der Zwischenabgangszeiten eines PH/PH/1/∞*–Systems (Abschnitt 3.2) folgt für den Vektor* τ_0'' *aus der Gleichung (3.28)* $\tau_0'' = 0$*, da das System fast nie „leer läuft“ und für den Vektor* τ_1'' *aus der Gleichung (3.29) die Beziehung* $\tau_1'' = \beta$*. Da* τ_1'' *nur vom Bedienprozeß und von der Populationsverteilung abhängt, ist offensichtlich im Fall* $\mu \to \lambda$ *die Populationsverteilung unabhängig vom zweiten Moment des Ankunftsprozesses. Der Erwartungswert der Population konvergiert demnach im Fall* $\mu \to \lambda$ *gegen unendlich, da dies insbesondere für den M/GI/1/*∞ *Fall gilt. Nach Little's Gesetz konvergiert also auch der Erwartungswert der Durchlaufzeit gegen unendlich. Da der Erwartungswert der Population gegen unendlich konvergiert, gilt dieselbe Aussage auch für die Varianz der Durchlaufzeitverteilung. Zusammenfassend gilt:*

$$\lim_{x \to \lambda} F_1(x, c_B) = \infty \tag{6.7}$$

$$\lim_{x \to \lambda} F_2(x, c_B) = \infty \tag{6.8}$$

□

Ebenso wie im Fall der strengen Monotonie der Funktionen F_1 und F_2 bei festem Variationskoeffizienten werden im folgenden Monotonieaussagen für den Fall fester Bedienratenraten gezeigt.

Lemma 6.2 *Für beliebige, fest gewählte $x = \mu \in]\lambda, \infty[$ sind die Funktionen*

1. $F_1 : \{(\mu, y) : y \in]0, \infty[\} \rightarrow]L_1(\mu), \infty[$
2. $F_2 : \{(\mu, y) : y \in]0, \infty[\} \rightarrow]L_2(\mu), \infty[$

stetig und streng monoton steigend. Dabei sind $L_1(\mu)$ und $L_2(\mu)$ gemäß $L_1(\mu) = \lim_{c \to 0} F_1(\mu, c)$ und $L_2(\mu) = \lim_{c \to 0} F_2(\mu, c)$ definiert. $L_1(\mu)$ bzw. $L_2(\mu)$ geben bei fest gewählter Bedienrate μ den Erwartungswert bzw. die Varianz der Durchlaufzeit eines $\mathcal{H}/\mathcal{H}/1$–Systems für den Fall an, daß der Variationskoeffizient der Bedienzeitverteilung gegen 0 konvergiert.

<u>*Beweis:*</u>

1. *Zum Beweis der strengen Monotonie seien die Funktionen $F_1(\mu, y)$ und $F_2(\mu, y)$ wiederum zunächst durch die Summe aus Erwartungswert bzw. Varianz der Bedienzeitverteilung und der Wartezeitverteilung $W(\mu, y)$ dargestellt, d.h.*

$$F_1(\mu, y) = \frac{1}{\mu} + E[W(\mu, y)] \tag{6.9}$$

$$F_2(\mu, y) = \frac{y^2}{\mu^2} + Var\left(W(\mu, y)\right) \tag{6.10}$$

   Sind c_1 und c_2 mit $0 < c_1 < c_2 < \infty$ die Variationskoeffizienten der Bedienzeitverteilungen B_1 und B_2 mit identischen Raten μ eines $\mathcal{H}/\mathcal{H}/1/\infty$ Systems, so gilt mit Satz A.10 $B_1 \overset{(2)}{\leq} B_2$. Im Fall $0 < c_1 < c_2 < 1$ ist zusätzlich zu fordern, daß die Repräsentationen von B_1 und B_2 gleiche Phasenlängen besitzen. Mit den Sätzen A.1 und A.2 und der Folgerung A.2 gilt für die Erwartungswerte und Varianzen der Wartezeitverteilungen $E[W(\mu, c_1)] < E[W(\mu, c_2)]$ und $Var(W(\mu, c_1)) < Var(W(\mu, c_2))$. Zusammenfassend ergibt sich bei festem μ die strenge Monotonie von F_1 und F_2 aus:

$$F_1(\mu, c_1) = \frac{1}{\mu} + E[W(\mu, c_1)] < \frac{1}{\mu} + E[W(\mu, c_2)] = F_1(\mu, c_2)$$

   und

$$F_2(\mu, c_1) = \frac{c_1^2}{\mu^2} + Var\left(W(\mu, c_1)\right) < \frac{c_2^2}{\mu^2} + Var\left(W(\mu, c_2)\right) = F_2(\mu, c_2)$$

2. *Unter der Voraussetzung, daß $E[W(\mu, y)]$ und $Var\left(W(\mu, x)\right)$ für festes $\mu \in]\lambda, \infty[$ auf dem Intervall $y \in]0, \infty[$ stetig sind, sind $F_1(\mu, y)$ und $F_2(\mu, y)$ als Summe stetiger Funktionen (vgl. (6.9)) ebenfalls auf dem Intervall $y \in]0, \infty[$ stetig. Auf die Stetigkeit von $E[W(\mu, y)]$ und $Var\left(W(\mu, x)\right)$ wird an dieser Stelle nicht weiter eingegangen.*

3. *Die Bildbereiche der Funktionen $F_1(x, y)$ und $F_2(x, y)$ für festes $\mu \in]\lambda, \infty[$ ergeben sich aus den folgenden Überlegungen:*

   Konvergiert y gegen unendlich, so konvergiert auch die Varianz der Bedienzeitverteilung gegen unendlich. Da sich die Varianz der Durchlaufzeitverteilung aus der Summe

der Varianzen der Bedienzeit und der Wartezeit ergibt, konvergiert diese offensichtlich ebenfalls gegen unendlich, d.h.

$$\lim_{x\to\infty} F_2(\mu, y) = \infty.$$

Mit der Krämer/Langenbach–Belz–Approximation (vgl. Gleichung 2.6) konvergiert auch der Erwartungswert der Durchlaufzeit gegen unendlich.

Anmerkung: Die Stetigkeit und strenge Monotonie der Funktionen $F_1(\mu, y)$ und $F_2(\mu, y)$ im Punkt $y = 1$ ergeben sich leicht aus der Tatsache, daß sowohl die $\mathcal{H}^{1-}$– als auch die $\mathcal{H}^{1+}$–Verteilung mit dem Variationskoeffizienten c_B für $c_B \to 1$ gegen die negative Exponentialverteilung konvergieren.

□

Mit den Lemmata 6.1 und 6.2 ist bereits eine wichtige Basis für das Iterationsverfahren gelegt. Die folgenden Sätze zeigen, wie ausgehend von einem Iterationswert μ_i bzw. c_i der jeweils nächste Wert unter Berücksichtigung einer der Bedingungen (6.1) bzw. (6.2) ermittelt wird. Dazu wird im folgenden zunächst der Begriff der Niveaulinie oder Höhenlinie mehrdimensionaler Funktionen definiert.

Definition 6.2 *Sei $F : \mathbb{R}^n \to \mathbb{R}$ eine n–dimensionale Funktion. Dann heißt $F^{-1}(y) := \{(x_1, \ldots, x_n) : F((x_1, \ldots, x_n) = y\}$ Niveaulinie oder Höhenlinie von F zum Niveau y.*

Im Fall einer zweidimensionalen Funktion läßt sich diese Situation graphisch veranschaulichen. In der Abbildung 6.2 ist ein Ausschnitt der Funktion $F_1(x, y)$ (für den Fall $\lambda = 1$ und $c_A = 1$) dargestellt. Ferner ist die Ebene $\tilde{F}(x, y) = 10$ eingezeichnet. Die Niveaulinie von F_1 zum Niveau 10 ergibt sich aus dem Schnitt der Funktion F_1 mit der Ebene $\tilde{F}$ bzw. aus der Projektion dieses Schnitts auf das durch x und y aufgespannte Koordinatensystem.

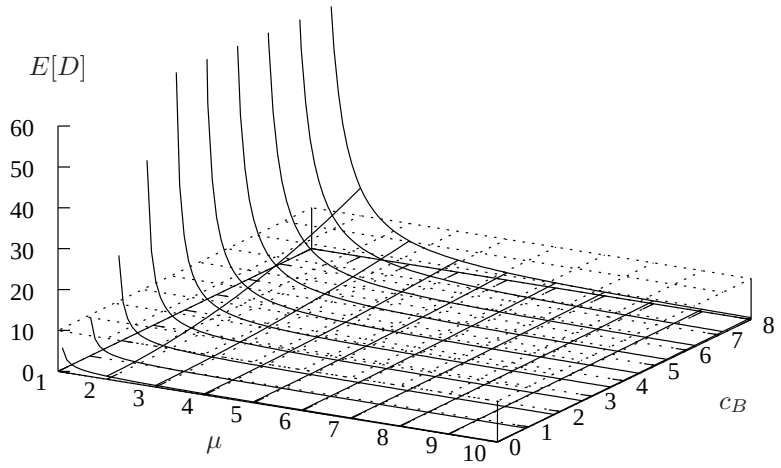

Abbildung 6.2: Konstruktion von M_{ED}

Das Ziel dieses Abschnittes, nämlich die Bestimmung von Parametern (μ, c_B), die die Gleichungen (6.1) und (6.2) erfüllen, ist mit der Definition 6.2 äquivalent zu der Bestimmung der Schnittmenge von $F_1^{-1}(ED)$ und $F_2^{-1}(VD)$. Zur Bestimmung von Elementen aus dieser Schnittmenge wird im folgenden zunächst auf Basis von $F_1^{-1}(ED)$ die Funktion M_{ED} definiert, die bei Kenntnis eines Iterationswertes c_i die Bedingung 6.1 herstellt.

Lemma 6.3 *Sei $ED \in]0, \infty[$ beliebig, fest gewählt. Dann ist die durch*

$$M_{ED}(y) = x \Leftrightarrow (x, y) \in F_1^{-1}(ED) = \{(x, y) : F_1(x, y) = ED\}.$$

definierte Funktion $M_{ED}(y)$ für alle $y \in]0, \infty[$ wohldefiniert. Ferner gilt:

$$M_{ED} :]0, \infty[\rightarrow]L_M(ED), \infty[$$

ist stetig und streng monoton steigend. Dabei gilt:

$$L_M(ED) = \mu_0 \Leftrightarrow \lim_{c_B \rightarrow 0} F_1(\mu_0, c_B) = ED. \tag{6.11}$$

<u>*Beweis:*</u>

1. *Die Wohldefiniertheit der Funktion $M_{ED}(y)$ für alle $y \in]0, \infty[$ folgt unmittelbar aus der Stetigkeit und der strengen Monotonie der Funktion $F_1(x, y)$ für jeweils feste Werte y.*

2. *Zum Beweis der strengen Monotonie von M_{ED} seien Werte $0 < c_1 < c_2 < \infty$ betrachtet. Dann gibt es gemäß Lemma 6.1 Werte μ_1 und μ_2 mit der Eigenschaft*

   $$ED = F_1(\mu_1, c_1) = F_2(\mu_2, c_2).$$

   Aus der strengen Monotonie der Funktion F_1 bei festem $x = \mu_2$ folgt:

   $$F_1(\mu_1, c_1) = F_1(\mu_2, c_2) > F_1(\mu_2, c_1)$$

   Daraus und aus der strengen Monotonie von F_1 bei festem $y = c_1$ folgt $\mu_1 < \mu_2$ und mit

   $$M_{ED}(c_1) = \mu_1 < \mu_2 = M_{ED}(c_2)$$

   die strenge Monotonie von M_{ED}.

3. *Der Bildbereich von M_{ED} ist nach unten durch $L_M(ED)$ mit der Eigenschaft*

   $$L_M(ED) = \min\{x : (x, y) \in F_1^{-1}(ED)\}$$

   beschränkt. Aufgrund der strengen Monotonie von M_{ED} ist dies äquivalent zu Gleichung 6.11. Ferner ist der Bildbereich von M_{ED} nach oben unbeschränkt. Gäbe es andererseits einen endlichen Wert μ mit $M_{ED}(y) \leq \mu$ für alle $y \in]0, \infty[$, so folgte aus der strengen Monotonie von M_{ED} die Bedingung $\lim_{y \rightarrow \infty} M_{ED}(y) = \mu$. Daraus folgte weiterhin mit Lemma 6.1 $ED = \lim_{y \rightarrow \infty} F_1(\mu, y) = \infty$ im Widerspruch zur Endlichkeit von ED.

4. Zum Beweis der Stetigkeit der Funktion M_{ED} auf dem gesamten Definitionsbereich sei eine beliebige Stelle $y_0 \in]0,\infty[$ mit $M_{ED}(y_0) = \mu_0$ betrachtet. Wäre M_{ED} in y_0 unstetig, so gäbe es einen reellen Wert $0 < \epsilon < \infty$ mit

$$\lim_{y \searrow y_0} M_{ED}(y) = \mu_0 + \epsilon \tag{6.12}$$

oder

$$\lim_{y \nearrow y_0} M_{ED}(y) = \mu_0 - \epsilon \tag{6.13}$$

Da beide Fälle auf gleiche Weise behandelbar sind, wird im folgenden ausschließlich der Fall 6.12 betrachtet. Gleichung 6.12 ist äquivalent zu

$$ED = F_1(\mu_0, y_0) = \lim_{y \searrow y_0} F_1(\mu_0 + \epsilon, y).$$

Sei weiter ein Wert μ_1 mit $\mu_0 < \mu_1 < \mu_0 + \epsilon$ betrachtet. Dann folgt aus der strengen Monotonie der Funktion $F_1(x, y)$ für festes $y = y_0$:

$$F_1(\mu_1, y_0) < F_1(\mu_0, y_0) = ED$$

und

$$\lim_{y \searrow y_0} F_1(\mu_1, y) > \lim_{y \searrow y_0} F_1(\mu_0 + \epsilon, y) = ED$$

Zusammenfassend ergibt sich:

$$F_1(\mu_1, y_0) < ED < \lim_{y \searrow y_0} F_1(\mu_1, y).$$

Hieraus folgte die Unstetigkeit der Funktion $F_1(x, y)$ für festes $x = \mu_1$ an der Stelle $y = y_0$ im Widerspruch zu Lemma 6.2. Folglich ist M_{ED} auf dem gesamten Definitionsbereich stetig.

□

Wie im Fall von $F_1^{-1}(ED)$ wird im folgenden auf Basis von $F_2^{-1}(VD)$ die Funktion $\tilde{M}_{VD}$ definiert.

Lemma 6.4 *Sei $VD \in]0,\infty[$ beliebig, fest gewählt. Dann ist die durch*

$$\tilde{M}_{VD}(y) = x \Leftrightarrow (x, y) \in F_2^{-1}(VD) = \{(x, y) : F_2(x, y) = VD\}$$

definierte Funktion $\tilde{M}_{VD}(y)$ für alle $y \in]0,\infty[$ wohldefiniert. Ferner gilt:

$$\tilde{M}_{VD} :]0,\infty[\rightarrow]L_{\tilde{M}}(VD), \infty[$$

ist stetig und streng monoton steigend mit

$$L_{\tilde{M}}(VD) = \mu_0 \Leftrightarrow \lim_{c_B \rightarrow 0} F_2(\mu_0, c_B) = VD. \tag{6.14}$$

Beweis:

1. *Die Wohldefiniertheit der Funktion $\tilde{M}_{VD}(y)$ für alle $y \in]0,\infty[$ folgt unmittelbar aus der Stetigkeit und der strengen Monotonie der Funktion $F_2(x,y)$ für jeweils feste Werte y.*

2. *Zum Nachweis der strengen Monotonie seien Werte $0 < c_1 < c_2 < \infty$ betrachtet. Dann existieren gemäß Lemma 6.1 Werte $\mu_1, \mu_2 > \lambda$ mit der Eigenschaft*

$$VD = F_2(\mu_1, c_1) = F_2(\mu_2, c_2).$$

Aus der strengen Monotonie von $F_2(x,y)$ bei festem $x = \mu_2$ folgt:

$$F_2(\mu_1, c_1) = F_2(\mu_2, c_2) > F_2(\mu_2, c_1).$$

Weiter folgt aus der strengen Monotonie von $F_2(x,y)$ bei festem $y = c_1$ die Bedingung $\mu_1 < \mu_2$ und aus

$$\tilde{M}_{VD}(c_1) = \mu_1 < \mu_2 = \tilde{M}_{VD}(c_2)$$

die strenge Monotonie von $\tilde{M}_{VD}$.

3. *Der Bildbereich von $\tilde{M}_{VD}$ ist nach unten durch $L_{\tilde{M}}(VD)$ mit der Eigenschaft*

$$L_{\tilde{M}}(VD) = \min\{x : (x,y) \in F_2^{-1}(VD)\}$$

beschränkt. Aufgrund der strengen Monotonie von $\tilde{M}_{VD}$ ist dies äquivalent zu Gleichung 6.14. Ferner ist der Bildbereich von $\tilde{M}_{VD}$ nach oben unbeschränkt.
Gäbe es andererseits einen endlichen Wert μ mit $\tilde{M}_{VD}(y) \leq \mu$ für alle $y \in]0,\infty[$, so folgte insbesondere aus der strengen Monotonie von $\tilde{M}_{VD}$ $\lim\limits_{y\to\infty} \tilde{M}_{VD}(y) = \mu$. Daraus folgt weiter mit Lemma 6.2: $VD = \lim\limits_{y\to\infty} F_2(\mu, y) = \infty$ im Widerspruch zur Endlichkeit von VD.

4. *Die Stetigkeit von $\tilde{M}_{VD}$ läßt sich ebenso zeigen, wie die Stetigkeit der Funktion M_{ED} in Lemma 6.3.*

□

Da die Funktionen M_{ED} und $\tilde{M}_{VD}$ stetig und streng monoton sind, existieren die Umkehrabbildungen. Diese werden in den folgenden Sätzen behandelt.

Lemma 6.5 *Sei $ED \in]0,\infty[$ beliebig, fest gewählt. Dann ist die durch*

$$\tilde{Z}_{ED}(x) = y \Leftrightarrow (x,y) \in F_1^{-1}(ED) = \{(x,y) : F_1(x,y) = ED\}$$

definierte Funktion $\tilde{Z}_{ED}$ wohldefiniert. Ferner gilt:

$$\tilde{Z}_{ED} :]L_M(ED), \infty[\to]0,\infty[$$

ist stetig und streng monoton steigend.

Beweis:

Der Beweis folgt unmittelbar aus der Tatsache, daß $\tilde{Z}_{ED}$ und M_{ED} invers zueinander sind.

□

Lemma 6.6 *Sei $VD \in]0,\infty[$ beliebig, fest gewählt. Dann ist die durch*

$$Z_{VD}(x) = y \Leftrightarrow (x,y) \in F_2^{-1}(VD) = \{(x,y) : F_2(x,y) = VD\}$$

definierte Funktion Z_{VD} wohldefiniert. Ferner gilt:

$$Z_{VD} :]L_{\tilde{M}}(VD), \infty[\rightarrow]0,\infty[$$

ist stetig und streng monoton steigend.

Beweis:

Die Aussage folgt unmittelbar aus der Tatsache, daß Z_{VD} und $\tilde{M}_{VD}$ invers zueinander sind.

□

Praktisch lassen sich die Funktionen M_{ED}, $\tilde{M}_{VD}$, $\tilde{Z}_{ED}$ und Z_{VD} durch ein beliebiges Verfahren zur Berechnung von Nullstellen errechnen. In obigen Ausführungen wurde bereits erwähnt, daß die Parameter der in diesem Abschnitt gesuchten Aggregate in der Schnittmenge $F_1^{-1}(ED) \cap F_2^{-1}(VD)$ liegen. Mit der Definition der Funktionen M_{ED} und $\tilde{M}_{VD}$ gilt für Parameterpaare $(\mu, c_B) \in F_1^{-1}(ED) \cap F_2^{-1}(VD)$ die Bedingung

$$\mu = M_{ED}(c_B) = \tilde{M}_{VD}(c_B).$$

Da $\tilde{M}_{VD}$ und Z_{VD} invers zueinander sind, d.h. $\mu = \tilde{M}_{VD}(c_B) \Leftrightarrow c_B = Z_{VD}(\mu)$, sind die Raten μ der gesuchten Aggregate Fixpunkte der im folgenden Satz definierten Funktion $G_{ED,VD}$.

Lemma 6.7 *Definiere für feste $\lambda, c_A \in]0,\infty[$, beliebige, feste Werte $ED \in]0,\infty[$ und $VD \in]0,\infty[$ die Funktion*

$$G_{ED,VD}(x) := M_{ED}(Z_{VD}(x)).$$

Dann gilt:

$$G_{ED,VD} :]L_{\tilde{M}}(VD), \infty[\rightarrow]L_M(ED), \infty[$$

ist stetig und streng monoton steigend.

Beweis:

Der Beweis folgt unmittelbar aus den Lemmata 6.3 und 6.6.

□

Da andererseits M_{ED} und $\tilde{Z}_{ED}$ invers zueinander sind, gilt für jeden Fixpunkt μ der Funktion $G_{ED,VD}$ die Bedingung $\tilde{Z}_{ED}(\mu) = Z_{VD}(\mu)$, und $(\mu, \tilde{Z}_{ED}(\mu)) = (\mu, Z_{VD}(\mu))$ liegt in der

Schnittmenge $F_1^{-1}(ED) \cap F_2^{-1}(VD)$. Somit existieren Parameter (μ, c_B), die die Gleichungen (6.1) und (6.2) erfüllen genau dann, wenn $G_{ED,VD}$ mindestens einen Fixpunkt besitzt bzw. wenn die Menge $F_1^{-1}(ED) \cap F_2^{-1}(VD)$ nicht leer ist. Eine hinreichende Bedingung für die Existenz eines Aggregats im Fall $VD \neq ED^2$ gibt das folgende Lemma an. Der Fall $VD = ED^2$ wird gesondert in Abschnitt 6.2 behandelt.

Lemma 6.8 *Seien $ED \in]0, \infty[$ und $VD \in]0, \infty[$ beliebig gewählt. Dann besitzt die Funktion $G_{ED,VD}$ mindestens einen Fixpunkt, wenn es im Fall $VD < ED^2$ bzw. ($VD > ED^2$) einen Wert $0 < c' < 1$ bzw. $1 < c' < \infty$ gibt, der die Bedingung*

$$F_1(\tilde{M}_{VD}(c'), c') \geq ED \text{ bzw. } F_1(\tilde{M}_{VD}(c'), c') \leq ED \tag{6.15}$$

erfüllt.

Beweis:

In Abschnitt 6.2 wird gezeigt, daß für alle $\mu > \lambda$ die Beziehung $(F_1(\mu, 1))^2 = F_2(\mu, 1)$ gilt (vgl. Gl. (6.28)). Somit folgt im Fall $VD < ED^2$ bzw. $VD > ED^2$

$$VD = F_2(\tilde{M}_{VD}(1), 1) \quad = \quad (F_1(\tilde{M}_{VD}(1), 1))^2 < ED^2 \text{ bzw.} \tag{6.16}$$

$$VD = F_2(\tilde{M}_{VD}(1), 1) \quad = \quad (F_1(\tilde{M}_{VD}(1), 1))^2 > ED^2. \tag{6.17}$$

Mit der Voraussetzung des Lemmas gilt im Fall $VD < ED^2$

$$F_1(\tilde{M}_{VD}(c'), c') \geq ED > F_1(\tilde{M}_{VD}(1), 1) \tag{6.18}$$

und im Fall $VD > ED^2$

$$F_1(\tilde{M}_{VD}(1), 1) > ED \geq F_1(\tilde{M}_{VD}(c'), c'). \tag{6.19}$$

Da die Funktion $F_1(x, y)$ in x bzw. y bei festem y bzw. x stetig ist, und da $\tilde{M}_{VD}(x)$ stetig ist, ist $F_1(\tilde{M}_{VD}(c), c)$ ebenfalls stetig. Somit existiert in beiden oben dargestellten Fällen ein Wert c mit $F_1(\tilde{M}_{VD}(c), c) = ED$, und $(\tilde{M}_{VD}(c), c)$ liegt offensichtlich in der Schnittmenge von $F_1^{-1}(ED)$ und $F_2^{-1}(VD)$. Somit ist $\tilde{M}_{VD}(c)$ Fixpunkt der Funktion $G_{ED,VD}$.

□

Wird in der Voraussetzung des Lemmas 6.8 die Einschränkung von c' auf die Intervalle $]0, 1[$ bzw. $]1, \infty[$ ausgelassen, so liefert Lemma 6.8 offensichtlich eine hinreichende und notwendige Bedingung für die Existenz eines Fixpunktes der Funktion $G_{ED,VD}$. Daß diese Bedingung dennoch hinzugenommen wird resultiert daraus, daß im Fall negativ–exponentiell verteilter Zwischenankunftszeiten die Funktion $F_1(\tilde{M}_{VD}(c), c)$ streng monoton fällt und zudem für $c \to \infty$ gegen 0 konvergiert. Diese Aussage wird in Abschnitt 6.3 in Lemma 6.9 gezeigt. Die Annahme ist daher die, daß diese Aussage auch in dem hier vorliegenden Fall $\mathcal{H}$ verteilter Zwischenankunftszeiten erhalten bleibt. Dies konnte im Rahmen dieser Arbeit jedoch nicht endgültig gezeigt werden. Trifft diese Annahme zu, so liefert Lemma 6.8 eine notwendige und hinreichende Bedingung für die Existenz eines Fixpunktes, die zudem im Fall $VD > ED^2$

stets gesichert ist, da die Bedingung $0 = \lim_{c' \to \infty} F_1(\tilde{M}_{VD}(c'), c') \leq ED$ stets erfüllt ist. Ferner ist unter der Annahme der strengen Monotonie der Funktion $F_1(\tilde{M}_{VD}(c), c)$ die Eindeutigkeit eines Fixpunktes gegeben. Hinsichtlich der Existenz eines Fixpunktes bleibt somit die Bedingung $\lim_{c \to 0} F_1(\tilde{M}_{VD}(c), c) \geq ED$ übrig.

Die Monotonieannahme der Funktion $F_1(\tilde{M}_{VD}(c), c)$ legt also nahe, die Überprüfung der Existenzbedingung (6.15) im Fall $VD < ED^2$ für einen sehr kleinen Variationskoeffizienten und im Fall $VD > ED^2$ für einen sehr großen Variationskoeffizienten durchzuführen. Da jedoch die Anzahl der Phasen einer $\mathcal{H}^{1-}$–Verteilung mit beliebig kleinem Variationskoeffizienten sehr groß wird (vgl. Anhang A.4), ist die Phasenanzahl einer $\mathcal{H}^{1-}$–Verteilung aus praktischer Sicht durch einen Wert K zu begrenzen. Auf diese Weise lassen sich Verteilungen mit einem minimalen Variationskoeffizienten $c^2_{min} = \frac{1}{K}$ approximieren. Andererseits ruft die Approximation von Verteilungen mit beliebig großen Variationskoeffizienten numerische Instabilitäten in den durchzuführenden Berechnungen hervor, so daß es aus praktischer Sicht ebenso sinnvoll ist, einen maximal erlaubten Variationskoeffizienten $1 < c_{max} < \infty$ zu wählen.

Daß es Fälle gibt, in denen die Bedingung $F_1(\tilde{M}_{VD}(c_{min}), c_{min}) \geq ED$ nicht erfüllbar ist, wird ebenfalls in Abschnitt 6.3 in Lemma 6.9 gezeigt. Im Fall der Existenz eines Fixpunktes liefert der folgende Satz eine Iterationsvorschrift, die diesen berechnet.

Satz 6.1 *Sind die Voraussetzungen des Lemmas 6.8 mit $c' = c_{min}$ im Fall $VD < ED^2$ und $c' = c_{max}$ im Fall $VD > ED^2$ erfüllt, so konvergiert die Folge $x_{i+1} = G_{ED,VD}(x_i)$ mit $x_0 = \tilde{M}_{VD}(c')$ gegen einen Fixpunkt der Funktion $G_{ED,VD}(x)$.*

Beweis:

Im folgenden wird lediglich der Fall $VD < ED^2$ betrachtet, da der Beweis des Falls $VD > ED^2$ in völlig analoger Weise durchführbar ist.

Aus dem Beweis des Lemmas 6.8 und aus der strengen Monotonie der Funktion $\tilde{M}_{VD}$ folgt die Existenz eines Fixpunktes $\mu \geq x_0$.

Ferner folgt aus der Voraussetzung des Lemmas 6.8:

$$F_1(\tilde{M}_{VD}(c'), c') \geq ED = F_1(M_{ED}(c'), c')). \tag{6.20}$$

Mit Lemma 6.2 gilt weiter $\tilde{M}_{VD}(c') \leq M_{ED}(c')$. Da Z_{VD} und $\tilde{M}_{VD}$ invers zueinander sind, folgt weiter:

$$x_1 = G_{ED,VD}(x_0) = M_{ED}(Z_{VD}(\tilde{M}_{VD}(c'))) = M_{ED}(c') \geq \tilde{M}_{VD}(c') = x_0. \tag{6.21}$$

Da $G_{ED,VD}$ eine streng monoton steigende Funktion ist, ist folglich $x_{i+1} = G_{ED,VD}(x_i)$ mit $x_0 = M_{ED}(c')$ eine monoton steigende Zahlenfolge. Da es ferner einen Fixpunkt $\mu \geq x_0$ gibt, ist die Folge zudem nach oben beschränkt und konvergiert somit gegen einen Fixpunkt.

□

Aus praktischer Sicht wird in Fällen, in denen die Voraussetzungen des Lemmas 6.8 für c_{min} bzw. c_{max} nicht erfüllt sind, keine weitere Berechnung durchgeführt. Alternativ zu dieser Vorgehensweise ließe sich das Aggregat mit den Parametern $(M_{ED}(c_{min}), c_{min})$ bzw. $(M_{ED}(c_{max}), c_{max})$ ermitteln. Dieses Aggregat erfüllt Gleichung (6.1) und liefert unter der Monotonieannahme der Funktion $F_1(\tilde{M}_{VD}(c), c)$ eine minimale Differenz $|VD - F_2(M_{ED}(c), c)|$.

Mit diesen Ausführungen läßt sich die Aussage dieses Abschnittes folgendermaßen zusammenfassen: Seien (λ, c_A) die Parameter des Ankunftsprozesses eines $\mathcal{H}/\mathcal{H}/1/\infty$ Systems, dessen Durchlaufzeit den Erwartungswert ED und die Varianz VD besitzt. Dann lassen sich unter den Voraussetzungen des Lemmas 6.8 Parameter (μ, c_B) mittels obiger Iterationsvorschrift finden, so daß Gleichungen (6.1) und (6.2) erfüllt sind.

In den folgenden beiden Abschnitten werden zwei Spezialfälle betrachtet, in denen einerseits der Fall $VD = ED^2$ behandelt wird und andererseits negativ–exponentiell verteilte Zwischenankunftszeiten betrachtet werden. In diesen Fällen lassen sich einige der in diesem Abschnitt dargestellten Aussagen deutlich vereinfachen.

6.2 $\mathcal{H}$/M/1/∞–Aggregate

In diesem Abschnitt wird der Spezialfall $ED^2 = VD$ betrachtet. Im folgenden wird gezeigt, daß unter vorgegebenen $\mathcal{H}$–verteilten Zwischenankunftszeiten mit den Parametern (λ, c_A) stets ein $\mathcal{H}$/M/1/∞–Aggregat existiert, dessen Durchlaufzeitverteilung den Erwartungswert ED und die Varianz VD besitzt. Dazu sei zunächst der allgemeinere Fall eines GI/M/1/∞–Systems betrachtet. In diesem Fall ist die Anzahl N' anwesender Kunden im System zum Zeitpunkt einer Ankunft eines neuen Kunden geometrisch mit dem Parameter σ verteilt ist (vgl. [49]), d.h.:

$$P\{N' = n\} = (1 - \sigma)\sigma^n \tag{6.22}$$

Diese Aussage bleibt sogar für GI/M/m/∞-FCFS Systeme erhalten, die statt eines einzelnen Bedieners m Bediener besitzen, also gleichzeitig mehrere Kunden bedienen können. Im Falle eines einzelnen Bedieners hängt der Parameter σ von der Laplace–Transformierten $L_A(s)$ der Zwischenankunftszeitverteilung ab und ergibt sich aus der eindeutigen Lösung der Gleichung

$$\sigma = L_A(\mu(1 - \sigma)) \tag{6.23}$$

im Intervall $0 < \sigma < 1$. Aus Gleichung 6.22 läßt sich der Erwartungswert $E[N']$ der Kundenanzahl im System zum Zeitpunkt einer Ankunft eines neuen Kunden ableiten zu:

$$E[N'] = \frac{\sigma}{(1 - \sigma)} \tag{6.24}$$

Weiterhin ist aus Gleichung 6.22 und der Laplace–Transformierten $L_B(s)$ der Bedienzeitverteilung die Laplace–Transformierte $L_D(s)$ der Durchlaufzeitverteilung folgendermaßen be-

stimmbar:

$$L_D(s) = \sum_{n=0}^{\infty}(1-\sigma)\sigma^n(L_B(s))^{n+1} \tag{6.25}$$

$$= \frac{1-\sigma}{\sigma}\sum_{n=1}^{\infty}\left(\frac{\sigma\mu}{s+\mu}\right)^n \quad (0<\sigma<1 \text{ und } s\geq 0) \tag{6.26}$$

$$= \frac{(1-\sigma)\mu}{s+(1-\sigma)\mu} \tag{6.27}$$

Die Durchlaufzeit eines GI/M/1/∞–Systems ist also offensichtlich negativ–exponentiell mit dem Parameter $(1-\sigma)\mu$ verteilt. Folglich gilt:

$$F_1(\mu,1) = \frac{1}{\mu(1-\sigma)} = \sqrt{F_2(\mu,1)}. \tag{6.28}$$

Unter Kenntnis des Erwartungswertes $0 < ED < \infty$ resultiert der Wert σ aus den Gleichungen (6.23) und (6.28) zu:

$$\sigma = L_A(1/ED), \tag{6.29}$$

und die Funktion $M_{ED}(1)$ hat die Darstellung

$$M_{ED}(1) = \frac{1}{ED(1-L_A(1/ED))} = \tilde{M}_{VD}(1). \tag{6.30}$$

Aus $0 < L_A(1/ED) < 1$ folgt unmittelbar $1/ED < M_{ED}(1) < \infty$. Ferner gilt $M_{ED}(1) > \lambda$, denn andernfalls folgte aus Lemma 6.1 $ED = F_1(M_{ED}(1),1) = \infty$ im Widerspruch zur Endlichkeit von ED. Im Fall $0 < ED^2 = VD < \infty$ existiert folglich stets ein $\mathcal{H}$/M/1/∞–Aggregat, dessen Durchlaufzeitverteilung den Erwartungswert ED und die Varianz VD besitzt. Die Rate μ der negativ–exponentiell verteilten Bedienzeit hat den Wert $\mu = M_{ED}(1) = \tilde{M}_{VD}(1)$. Der konkrete Wert von μ ergibt sich schließlich aus einer geeigneten Approximation der Zwischenankunftszeiten durch $\mathcal{H}$–Verteilungen mit den durch A.28 und A.48 gegebenen Laplace–Transformierten.

Für den Fall ebenfalls negativ–exponentiell verteilter Zwischenankunftszeiten, d.h. $c_A = 1$ ergibt sich das bekannte Resultat für M/M/1–Systeme

$$\mu = \lambda + \frac{1}{E[D]} \tag{6.31}$$

6.3 M/$\mathcal{H}$/1/∞–Aggregate

In diesem Abschnitt wird der Spezialfall betrachtet, daß der Variationskoeffizient c_A der fest vorgegebenen Zwischenankunftszeitverteilung den Wert 1 besitzt. Gemäß der in dieser Arbeit angewandten Approximation durch Phasenverteilungen (vgl. Anhang A) läßt sich die Zwischenankunftszeitverteilung folglich durch eine negative Exponentialverteilung mit der Rate λ approximieren. In dem Spezialfall eines $M/\mathcal{H}/1/\infty$ –Systems besitzen die in den

Lemmata 6.1 und 6.2 definierten Funktionen F_1 und F_2 explizite Darstellungen. So ergibt sich der Erwartungswert der Durchlaufzeitverteilung aus der Pollaczek–Khinchin Formel für M/G/1/∞ Systeme und Little's Gesetz. Die Pollaczek–Khinchin Formel gibt die mittlere Anzahl $E[N]$ von Kunden im System an. Im Unterschied zu den in Abschnitt 6.2 betrachteten $\mathcal{H}/M/1/\infty$ Systemen ist diese Anzahl in $M/G/1/\infty$ Systemen vom Betrachtungszeitpunkt unabhängig. Es gilt:

$$E[N] = \rho + \rho^2 \frac{(1 + c_B^2)}{2(1 - \rho)} \tag{6.32}$$

Dabei ist $\rho = \lambda/\mu$ die Systemauslastung. Mit Little's Gesetz resultiert aus Gleichung 6.32 der Erwartungswert $F_1(\mu, c_B)$ der Durchlaufzeitverteilung zu

$$F_1(\mu, c_B) = \frac{2\mu + \lambda(c_B^2 - 1)}{2\mu(\mu - \lambda)}. \tag{6.33}$$

Aus Gleichung 6.33 läßt sich die in Lemma 6.3 definierte Funktion M_{ED} für einen fest vorgegebenen Erwartungswert ED der Durchlaufzeit explizit angeben.

$$M_{ED}(c_B) = \frac{1 + \lambda ED + \sqrt{\lambda^2 ED^2 + 2\lambda ED c_B^2 + 1}}{2ED} \tag{6.34}$$

Neben dem Resultat von Pollaczek–Khinchin ist für M/G/1/∞–Systeme zudem die Laplace–Transformierte $L_D(s)$ der Durchlaufzeitverteilung bekannt. Diese ergibt sich aus der Laplace–Transformierten $L_B(s)$ der Bedienzeitverteilung folgendermaßen (vgl. [49]):

$$L_D(s) = \frac{L_B(s)s(1 - \rho)}{s - \lambda + \lambda L_B(s)} \tag{6.35}$$

Aus 6.35 resultieren die k-ten Momente $E[D^k]$ der Durchlaufzeitverteilung zu

$$E[D^k] = \lim_{s \to 0} \left((-1)^k \frac{d^k}{ds^k} L_D(s) \right). \tag{6.36}$$

Insbesondere läßt sich damit die Varianz $F_2(\mu, c_B)$ der Durchlaufzeit ermitteln. Dazu muß jedoch zunächst die Bedienzeitverteilung durch eine geeignete Phasenverteilung approximiert werden, um $L_B(s)$ zu bestimmen.

Im Fall $c_B \geq 1$ wird die Bedienzeitverteilung durch eine $\mathcal{H}^{1+}$–Verteilung approximiert (vgl. Anhang A.3), und somit folgt für $F_2(\mu, c_B)$:

$$F_2(\mu, c_B) = (F_1(\mu, c_B))^2 + \frac{(\lambda c_B^2 + \mu)(c_B^2 - 1)}{\mu^2(\mu - \lambda)}. \tag{6.37}$$

Im Fall $0 < c_B < 1$ wird die Bedienzeitverteilung durch eine $\mathcal{H}^{1-}$–Verteilung approximiert (vgl. Anhang A.4). Im folgenden wird jedoch der weiter eingeschränkte Fall betrachtet, daß die Bedienzeitverteilung durch eine $\mathcal{H}^{1-}$–Verteilungen mit zwei aufeinanderfolgenden negativ–exponentiell verteilten Phasen approximiert wird. Auf diese Weise lassen sich Variationskoeffizienten c_B im Intervall $c_{min}^2 = \frac{1}{2} \leq c_B^2 < 1$ approximieren. Die nachfolgenden Ausführungen

lassen sich leicht auf den Fall von mehr als zwei Phasen erweitern. Aus Darstellungsgründen wird jedoch auf die Beschreibung dieses allgemeinen Falls verzichtet. Die Varianz der Durchlaufzeitverteilung hat damit im Fall $c_{min}^2 \leq c_B^2 < 1$ die Darstellung

$$F_2(\mu, c_B) = (F_1(\mu, c_B))^2 - \frac{1 - c_B^2}{\mu(\mu - \lambda)}. \tag{6.38}$$

Auf dieser Grundlage beantwortet der folgenden Satz die Frage hinsichtlich der Existenz und Eindeutigkeit eines $M/\mathcal{H}/1/\infty$ –Aggregats.

Lemma 6.9 *Zu vorgegebenen Werten $ED \in]0, \infty[$ und $VD \in]0, \infty[$ existieren Parameter (μ, c_B) mit $c_B \geq c_{min}$ eines $M/\mathcal{H}/1/\infty$ –Aggregats genau dann, wenn gilt:*

$$VD \geq F_2(M_{ED}(c_{min}), c_{min}). \tag{6.39}$$

Im Falle der Existenz eines Aggregats ist dieses zudem eindeutig.

Beweis:

Sei zunächst die Funktion $F_2(M_{ED}(c), c)$ auf dem Bereich $1 \leq c < \infty$ untersucht. Aus der Funktion $F_2(M_{ED}(c), c)$, auf deren explizite Darstellung aufgrund ihrer Komplexität an dieser Stelle verzichtet wird, läßt sich folgender Grenzwert erkennen:

$$\lim_{c \to \infty} F_2(M_{ED}(c), c) = \infty. \tag{6.40}$$

Ferner ist die Ableitung von $F_2(M_{ED}(c), c)$ auf dem Bereich $1 \leq c < \infty$ strikt positiv. Da aus Gleichung (6.37) unmittelbar $F_2(M_{ED}(1), 1) = (F_1(M_{ED}(1), 1))^2 = ED^2$ folgt, bildet also $F_2(M_{ED}(c), c)$ Werte $1 \leq c < \infty$ streng monoton steigend auf den Bildbereich $[ED^2, \infty[$ ab. Im Fall $VD \geq ED^2$ existieren folglich stets Parameter $(M_{ED}(c_B), c_B)$ mit $c_B \geq 1$ eines $M/\mathcal{H}/1/\infty$ –Aggregats, und die Parameter sind aufgrund der strengen Monotonie von $F_2(M_{ED}(c), c)$ zudem eindeutig.

Im Fall $c_{min} \leq c < 1$ läßt sich die strenge Monotonie der Funktion $F_2(M_{ED}(c), c)$ unmittelbar anhand der Gleichung (6.38) erkennen, da $M_{ED}(c)$ eine streng monoton steigende Funktion ist. Folglich nimmt $F_2(M_{ED}(c), c)$ für $c = c_{min}$ das Minimum an. Damit gilt für Werte $c_{min} \leq c < 1$:

$$F_2(M_{ED}(c), c) \geq F_2(M_{ED}(c_{min}), c_{min}) = ED^2 - \frac{2ED^2}{(1+\lambda ED+\sqrt{1+\lambda ED+\lambda^2 ED^2})(1-\lambda ED+\sqrt{1+\lambda ED+\lambda^2 ED^2})} > \frac{ED^2}{2}. \tag{6.41}$$

Die Beziehung $F_2(M_{ED}(c_{min}), c_{min}) > \frac{ED^2}{2}$ resultiert aus der Bildung des Grenzwertes von (6.41) für $\lambda \to 0$ und der Tatsache, daß F_2 in λ monoton steigt (vgl. hierzu Satz A.1). Da aus Gleichung (6.38) unmittelbar $F_2(M_{ED}(1), 1) = (F_1(M_{ED}(1), 1))^2 = ED^2$ folgt, bildet $F_2(M_{ED}(c), c)$ Werte $c_{min} \leq c < 1$ streng monoton steigend auf den Bildbereich $[F_2(M_{ED}(c_{min}), c_{min}), ED^2[$ ab.

Zusammenfassend bildet $F_2(M_{ED}(c), c)$ Werte $c_{min} \leq c < \infty$ streng monoton steigend auf den Bereich $[F_2(M_{ED}(c_{min}), c_{min}), \infty[$ ab. Damit folgt die Behauptung des Satzes. Ferner folgt mit

den Lemmata 6.1 und 6.2, daß $F_1(\tilde{M}_{VD}(c), c)$ streng monoton fällt und für $c \to \infty$ gegen 0 konvergiert.

Sind zur Approximation der Bedienzeitverteilungen mit Variationskoeffizienten $0 < c_B < 1$ $K > 2$ Phasen erlaubt, so folgt $c^2_{min} = \frac{1}{K}$, und (6.41) hat die Darstellung

$$F_2(M_{ED}(c), c) \geq F_2(M_{ED}(c_{min}), c_{min}) > \frac{ED^2}{K}. \tag{6.42}$$

Die Bedingung des Satzes hinsichtlich der Existenz und Eindeutigkeit eines Aggregats bleibt somit für $c^2_{min} = \frac{1}{K}$ erhalten.

□

Im folgenden werden nun explizite Darstellungen für die Parameter (μ, c_B) eines $M/\mathcal{H}/1/\infty$ –Aggregats ermittelt. Dazu wird zunächst aus den Gleichungen (6.37) und (6.38) und Gleichung (6.33) die in Lemma 6.6 definierte Funktion Z_{VD} angeben. Im Fall $VD \geq ED^2$ hat Z_{VD} die Darstellung

$$Z_{VD}(\mu) = \sqrt{\frac{2(\mu-\lambda)(\sqrt{\mu^2+\lambda^2+4\lambda\mu^3 VD - 3\lambda^2\mu^2 VD} - \mu) - \lambda^2}{\lambda(4\mu - 3\lambda)}}. \tag{6.43}$$

Im Fall $F_2(M_{ED}(c_{min}), c_{min}) \leq VD < ED^2$ ergibt sich die Funktion Z_{VD} zu

$$Z_{VD}(\mu) = \sqrt{\frac{2\mu\sqrt{\mu^2 - \lambda^2 + \lambda^2 VD(\mu-\lambda)^2} - 2\mu^2 + \lambda^2}{\lambda^2}}. \tag{6.44}$$

Durch Lösen der Gleichung $\mu = M_{ED}(Z_{VD}(\mu))$ ergeben sich schließlich die gesuchten Parameter (μ, c_S) des Aggregats. Im Fall $VD \geq ED^2$ ergibt sich:

$$\mu = \frac{6ED + \lambda VD + 3\lambda ED^2 + A}{8ED^2} \tag{6.45}$$

$$A = \sqrt{(9\lambda^2 ED^2 + 15\lambda^2 VD + 4\lambda ED)(ED^2 - VD) + (2ED + 4\lambda VD)^2} \tag{6.46}$$

$$c_B^2 = \frac{2ED\mu(\mu-\lambda) - 2\mu + \lambda}{\lambda} \tag{6.47}$$

Ferner ergibt sich im Fall $F_2(M_{ED}(c_{min}), c_{min}) \leq VD < ED^2$:

$$\mu = \frac{\lambda^2(ED^2 - VD) + 2(\lambda ED + 1)}{2ED + \lambda(ED^2 - VD)} \tag{6.48}$$

$$c_B^2 = \frac{4VD - \lambda^2(ED^2 - VD)^2}{2ED + \lambda(ED^2 - VD)} \tag{6.49}$$

6.4 Ziel–Modellklasse

Abschließend zu den Ausführungen des zweiten Teils dieser Arbeit wird in diesem Abschnitt auf die Struktur der Warteschlangennetze QN_i aus Abbildung 6.1 eingegangen. Es wird aufgezeigt, welche Klasse von Warteschlangennetzen der Aggregierung zugänglich ist. Dazu müssen sich letztendlich die ersten beiden Momente der Durchlaufzeitverteilung berechnen lassen. Im einfachsten Fall sind die ersten beiden Momente der Durchlaufzeit der QN_i bekannt. Diese Kenntnis kann aus gewissen Modellannahmen resultieren oder auch aus der Verwendung alternativer Analyseverfahren wie z.B. der Simulation. In diesen Fällen lassen sich die Aggregate direkt berechnen.

Sind die Momente der Durchlaufzeit nicht bekannt, so kann das in dieser Arbeit vorgestellte Dekompositionsverfahren zu deren Ermittlung herangezogen werden. Auf diese Wiese lassen sich aus rein technischer Sicht erweiterte Fork/Join–Warteschlangennetze mit einem eindeutigen Eintritts– und Austrittsknoten aggregieren. Diese können die in Abschnitt 3.3 aufgeführten Stationstypen und die in Kapitel 6 vorgestellten erweiterten Fork/Join–Stationen enthalten. In der im Rahmen dieser Arbeit entstandenen Implementierung des Dekompositionsverfahrens für erweiterte Fork/Join–Warteschlangennetze wurden konkret die folgenden Stationstypen unter Verwendung der in Anhang A vorgestellten speziellen $\mathcal{H}$ Phasenverteilungen realisiert:

- $\mathcal{H}/\mathcal{H}/1/\infty$–FCFS Stationen,
- $\mathcal{H}$/M/m–FCFS Stationen,
- $\mathcal{H}/\mathcal{H}/\infty$ Stationen,
- Fork/Join–Stationen mit $\mathcal{H}$–verteiltem Ankunftsprozeß und $\mathcal{H}$ verteilten Bedienprozessen.

Hinsichtlich der Aggregierung von erweiterten Fork/Join–Warteschlangennetzen unter Verwendung des Dekompositionsverfahrens zur Bestimmung der Momente der Durchlaufzeit sind jedoch zwei Aspekte zu berücksichtigen. Wie bereits im Kontext der Untersuchungen der azyklischen bzw. zyklischen erweiterten Fork/Join–Warteschlangennetze im Abschnitt 5.3 erläutert wurde, liegt die inhärente Problematik des Dekompositionsverfahrens in der Vernachlässigung von Abhängigkeiten zwischen den isoliert betrachteten Stationen. Dies führt insbesondere zu einer Verfälschung des zweiten Moments der Gesamtdurchlaufzeit zyklischer Modelle und auch großer azyklischer Modelle. Ein zweiter Aspekt, der an dieser Stelle lediglich genannt, jedoch nicht weiter ausgeführt wird, ist der Aspekt des Überholens von Kunden. In dem zu aggregierenden Warteschlangennetz mag das Überholen von Kunden z.B. in einer $\mathcal{H}/\mathcal{H}/\infty$ Stationen erlaubt bzw. möglich sein. In dem entsprechend konstruierten $\mathcal{H}/\mathcal{H}/1\infty$–FCFS Aggregat ist offensichtlich ein Überholen von Kunden aufgrund der FCFS–Bediendisziplin nicht möglich. Die Auswirkungen von Überholungen in Warteschlangennetzen werden z.B. in [95] dargestellt.

Aus rein praktischer Sicht besteht hinsichtlich der Aggregierung durch $\mathcal{H}/\mathcal{H}/1/\infty$–FCFS Stationen ein weiteres Problem. Dieses Problem tritt dann auf, wenn der Erwartungswert der

Durchlaufzeit erheblich größer ist, als die Zwischenankunftszeiten oder wenn der Variationskoeffizient der Durchlaufzeit extrem klein oder extrem groß ist. In diesen Fällen besitzt das Aggregat eine sehr hohe Auslastung bzw. ebenfalls einen extrem kleinen oder großen Variationskoeffizienten. Im allgemeinen ist die numerische Analyse von $\mathcal{H}/\mathcal{H}/1\infty$–FCFS Stationen mit derartigen Eigenschaften recht fehleranfällig und sehr zeitintensiv. Andererseits haben die Erfahrungen beim Experimentieren mit den Aggregaten gezeigt, daß deren Berechnungszeit stets deutlich geringer ist, als die Rechenzeit, die zur Analyse darauf basierender Upper–Bound Modelle benötigt wird.

Zuletzt sei an dieser Stelle nochmals die in Abschnitt 5.2.4 bereits angesprochene Problematik der mit wachsender Anzahl paralleler Bediener drastisch abnehmenden Effizienz des Analyseverfahrens für das Upper–Bound Modell. Anhand der Aggregierungstechnik läßt sich aus rein technischer Sicht die Analyse eines Upper-Bound Modells mit $N > 2$ parallelen Bedienern durch einen Divide-and-Conquer Ansatz auf die Analyse von maximal $N - 1$ Upper–Bound Modellen mit zwei parallelen Bedienern reduzieren. Dadurch läßt sich der sehr hohe Rechen– und Speicherplatzbedarf des Analyseverfahren für den Fall $N > 2$ erheblich reduzieren. Andererseits induziert diese Vorgehensweise eine zusätzliche Approximation, die Einfluß auf die Qualität der Analyseresultate hat. Die Effizienzsteigerung des Analyseverfahrens durch diesen Divide-and-Conquer Ansatz als auch die Auswirkungen auf die Approximationsgüte der Analyseresultate werden im dritten Teil dieser Arbeit anhand eines Anwendungsbeispiels in Abschnitt 7.1 näher beleuchtet.

Teil III

Anwendungsgebiete

Kapitel 7

Verteilte Computer- und Kommunikationssysteme

Dieser dritte Teil der Arbeit stellt den Anwendungsbezug der in den ersten beiden Teilen erarbeiteten Analysemethoden her. Aus dem vielfältigen Spektrum nebenläufiger Systeme werden verschiedene aktuelle Anwendungsfälle aus zwei Bereichen betrachtet.

Dieses Kapitel beschäftigt sich mit dem Gebiet Computer- und Kommunikationssysteme. Die Analyse verteilter bzw. paralleler Systeme in diesem Bereich ist seit langer Zeit ein wichtiges Thema in der Informatik. Desweiteren werden in Kapitel 8 Anwendungsfälle aus dem Bereich Logistik und Produktion betrachtet. Typische Anwendungsbeispiele aus dem Bereich Computer- und Kommunikationssysteme sind Multi–Prozessorsysteme, verteilte Kommunikationsprotokolle, verteilte Rechnernetzanwendungen, verteilte Datenbanken, RAID-Systeme, verteilte/parallele Simulation u.v.a. Diese Themen werden in der Informatik seit langer Zeit behandelt [14, 70, 10, 64, 39, 90].

Eine gegenüber diesen Themen recht junge Domäne ist die verteilter Web–basierter Informationsdienste. Aufgrund der stetig steigenden Verfügbarkeit und Akzeptanz des Internets gewinnen Web–Dienstleistungen zunehmend an Bedeutung. Einige typische Beispiele sind Meta–Suchmaschinen, E–Shops, Preisfinder, Flug- und Urlaubs–Buchungssysteme u.v.a. Vielen dieser Systeme ist ein hoher Grad an Parallelität gemein. Meta–Suchmaschinen wie *www.metacrawler.com*, *www.mamma.com* oder *www.metager.de* senden Suchanfragen parallel an mehrere (Basis-) Suchmaschinen und bereiten die Resultate geeignet als Antwort auf die Suchanfrage auf. Informationsdienste wie *www.traveloverland.de*, *www.opodo.de* und *www.expedia.de* bieten dem Benutzer die Möglichkeit, nach Flügen und Urlaubsreisen unterschiedlicher Anbieter zu suchen und diese anschließend online zu buchen. Auch in diesem Fall wird eine Suchanfrage parallel an die jeweiligen Informationssysteme der Fluggesellschaften und Reiseveranstalter weitergeleitet. Nach Eintreffen der Resultate erhält der Benutzer z.B. eine preislich sortierte Liste aller verfügbaren Flüge oder Urlaubsreisen in einem gewissen Zeitraum. Eine allgemeine Skizze all dieser Informationsdienste ist in der Abbildung 7.1 am Beispiel einer Meta–Suchmaschine skizziert.

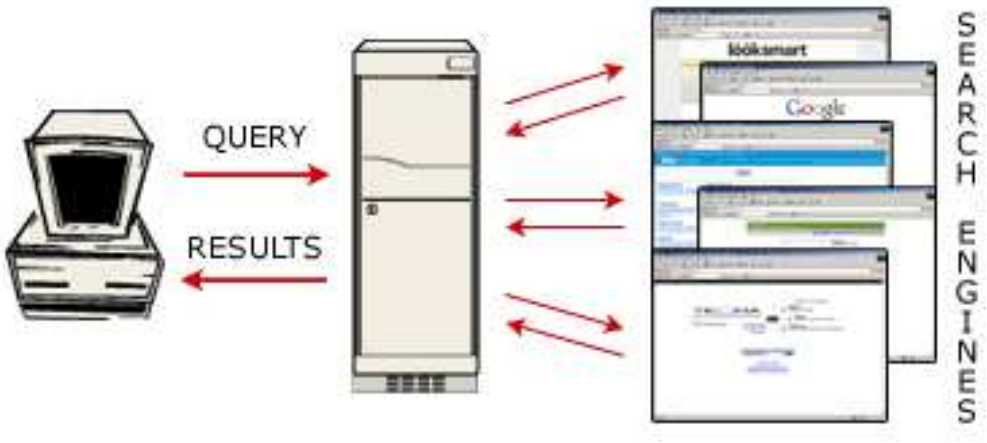

Abbildung 7.1: Arbeitsweise verteilter Web–basierter Informationsdienste (Quelle: www.mammo.com)

Die Akzeptanz der Informationsdienste hängt wesentlich von gewissen qualitativen und quantitativen Eigenschaften ab. Wichtige Kriterien sind die Antwortzeit und die Relevanz der Ergebnisse. Im folgenden Abschnitt werden Meta–Suchmaschinen genauer betrachtet.

7.1 Meta–Suchmaschinen

Das Internet hält riesige Mengen an Informationen zu nahezu jedem Themenbereich vor. Schätzungen gehen davon aus, daß Ende 2004 etwa 10-15 Mrd. Dokumente im Internet verfügbar waren und daß sich die Anzahl etwa halbjährlich verdoppelt. Um diese enorme Datenflut beherrschbar zu machen und schließlich nutzbringend einsetzen zu können, unterstützen Internet–Suchmaschinen wie z.B. Google, Altavista, Excite, Lycos u.v.a. den Benutzer darin, zu einem bestimmten Thema eine übersichtliche Anzahl relevanter Informationen bzw. Dokumente zu finden. Dabei müssen die Dienste unterschiedliche qualitative und quantitative Anforderungen erfüllen. Ein wichtiges Kriterium ist die Relevanz der Suchergebnisse, d.h. beantworten die gefundenen Dokumente auch tatsächlich die Suchanfrage des Benutzers. Ein weiteres Kriterium ist die Vollständigkeit der Suchergebnisse. Daran ist die Erwartung geknüpft, daß zu einer Suchanfrage alle „wichtigen“ Dokumente präsentiert werden und keine Informationslücken übrig bleiben. Ein drittes wichtiges Kriterium ist die Antwortzeit des Systems, da Benutzer im allgemeinen nicht bereit sind, beliebig lange auf eine Antwort zu warten.

Um diesen unterschiedlichen Anforderungen gerecht zu werden, besitzen Suchdienste eine recht komplexe Architektur. Eine wesentliche Komponente bildet eine Datenbank, in der strukturierte Informationen zu allen suchbaren, d.h. der Suchmaschine bekannten Dokumenten eingepflegt werden. Detaillierte Informationen dazu, wie z.B. Dokumente einem Ranking unterzogen werden, welche Methoden eingesetzt werden, um Übereinstimmungen zwischen einer Suchanfrage und Dokumenteninhalten festzustellen und auf welche Weise Web–Robots den Datenbankbestand einer Suchmaschine erweitern und aktualisieren, werden in [30, 82, 25]

gegeben.

Kritisch zu betrachten ist jedoch der Aspekt der Vollständigkeit. Aufgrund der enormen Anzahl im Internet verfügbarer Dokumente und der sehr großen Anzahl an Nutzern wird eine einzelne Suchmaschine kaum den Bestand des gesamten Internets abdecken bzw. durchsuchen können. Daher wurden Meta–Suchmaschinen mit dem Ziel entwickelt, durch die Nutzung mehrerer der soeben vorgestellten Suchmaschinen einen möglichst großen Teil der im Internet verfügbaren Dokumente abzudecken. Zur Abgrenzung der beiden verschiedenen Typen von Internet–Suchmaschinen werden im folgenden die Begriffe Meta–Suchmaschine und Basis–Suchmaschine unterschieden. Um auf die in der Informatik übliche Schichten–Architektur von Systemen hinzuweisen, wird alternativ auch der Begriff Basisdienst verwendet.

Im Unterschied zu Basis–Suchmaschinen pflegen Meta–Suchmaschinen keine eigene Datenbank mit Informationen über im Internet verfügbare Dokumente. Stattdessen besitzen sie gewisse Informationen über die von ihnen genutzten Basisdienste. Diese Informationen umfassen z.B. spezifische Anfrageformate und Anfrageoptionen, Ranking–Kriterien, falls die Basisdienste diese preisgeben und Themen–spezifische Prioritäten. Da sich die Anfrageformate und Anfrageoptionen verschiedener Basisdienste i.a. unterscheiden, wird die Suchanfrage eines Benutzers von der Meta–Suchmaschine zunächst in Basisdienst–spezifische Suchanfragen transformiert. Anschließend werden die Anfragen parallel an die von der Meta–Suchmaschine genutzten Basisdienste gesendet. Nach dem Eintreffen der Resultate werden diese gemäß eines eigenen Rankings sortiert, Duplikate werden verworfen, und letztlich wird dem Benutzer das Such–Ergebnis präsentiert.

Anhand der folgenden Abbildung 7.2 wird der schematische Aufbau einer typischen Meta–Suchmaschine (vgl. [81]) genauer veranschaulicht. Zunächst hat der Benutzer über eine Eingabemaske die Möglichkeit, eine Suchanfrage sowie gewisse Suchoptionen zu formulieren. Anschließend wird die Suchanfrage von dem Eingabeübersetzer in die spezifischen Eingabeformate der von der Meta–Suchmaschine genutzten Basis–Suchmaschinen transformiert. Teilweise erfolgt zusätzlich eine Selektion der genutzten Basisdienste, die z.B. anhand gewisser Schlüsselworte in der Suchanfrage Themen–spezifische Basisdienste einschließt und Basisdienste mit abweichendem Themenschwerpunkt aussortiert. Schließlich werden die transformierten Suchanfragen parallel an die ausgewählten Basis–Suchmaschinen gesendet. Ist das Resultat eines Basisdienstes in der Meta–Suchmaschine eingetroffen so ist evtl. ein Download der gelieferten Referenzen notwendig. Dies ist dann der Fall, wenn entweder keine Ranking–Informationen über die entsprechende Basis–Suchmaschine verfügbar sind oder wenn sich die Ranking–Kriterien erheblich von denen der Meta–Suchmaschine unterscheiden oder wenn die eingestellten Suchoptionen der Meta–Suchmaschine von dem Basisdienst nicht unterstützt werden. In den ersten beiden Fällen ist ein nachträgliches Ranking der Referenzen anhand der Downloads möglich, und im dritten Fall ermöglichen die Downloads eine Selektion der Referenzen hinsichtlich der eingestellten Suchoptionen in der Meta–Suchmaschine. Sind auf diese Weise die Resultate aller Basisdienste verfügbar, werden abschließend Duplikate verworfen und die übrigen Referenzen hinsichtlich des Rankings sortiert. Ggfls. bleiben in diesen Schritten die Resultate von Basisdiensten, die nach einem gewissen Time–out kein Ergebnis geliefert haben, unberücksichtigt. Letztendlich wird dem Benutzer das Suchergebnis gesendet.

Das Interesse dieses Abschnitts besteht in der quantitativen Analyse einer derartigen Meta–

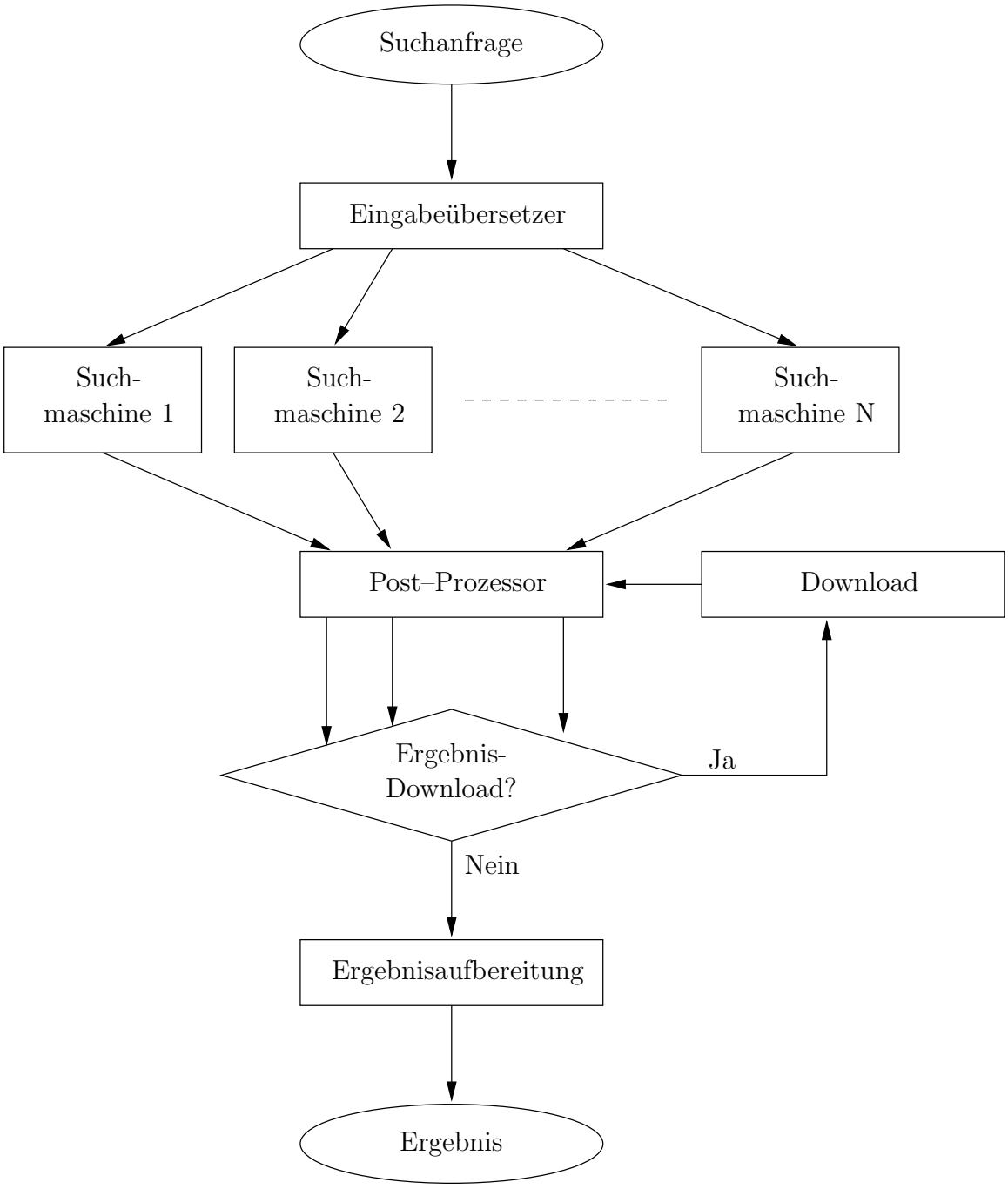

Abbildung 7.2: Aufbau einer Meta–Suchmaschine

Suchmaschine bzgl. ihrer Leistungsfähigkeit. Speziell wird die Antwortzeit betrachtet, d.h. die Zeitspanne zwischen dem Absenden einer Suchanfrage seitens eines Benutzers und der Ergebnispräsentation der Meta–Suchmaschine. Wie bereits erwähnt, hat die Antwortzeit wesentlichen Einfluß auf die Akzeptanz eines Web–basierten Informationsdienstes. Konkret werden zwei Fragestellungen betrachtet. In der Entwicklungsphase einer Meta–Suchmaschine wird untersucht, welche technische Rahmenbedingungen benötigt werden, ein erwartetes Lastspektrum hinsichtlich einer vorgegebenen Antwortzeit bewältigen zu können. Ferner wird in der Betriebsphase einer Meta–Suchmaschine untersucht, welche Leistungsfähigkeit bei einer zunehmenden Last zu erwarten ist.

7.1.1 Modellbeschreibung

Um dem obigen Fragestellungen nachzugehen, wird das in der Abbildung 7.3 dargestellte Warteschlangenmodell einer Meta–Suchmaschine betrachtet.

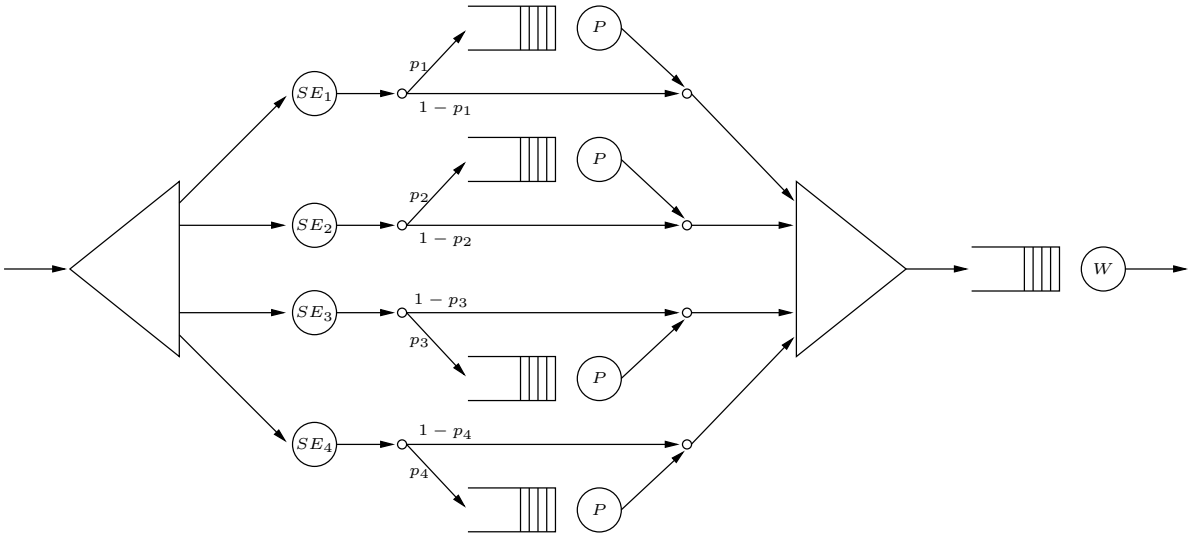

Abbildung 7.3: Warteschlangenmodell einer Meta–Suchmaschine

In dem Modell werden gegenüber der obigen detaillierten Beschreibung einer Meta–Suchmaschine einige Vereinfachungen angenommen. So bleibt der Eingabeübersetzer unberücksichtigt, da davon auszugehen ist, daß die Zeit zur Transformation der Suchanfragen nahezu vernachlässigbar ist. Ferner wird angenommen, daß die Meta–Suchmaschine stets dieselben vier Basis–Suchmaschinen nutzt und ein Time-out nicht auftritt.

Die vier Basis–Suchmaschinen sind durch die Stationen $SE_1, \ldots, SE_4$ dargestellt. Im allgemeinen hat der Planer der Meta–Suchmaschine keine exakten Informationen über die technische und technologische Ausprägung der Basisdienste. Er kann jedoch aufgrund von Informationen des Basisdienst–Betreibers oder durch eigene Messungen die Antwortzeit der Basisdienste ermitteln. Daher sind die Suchmaschinen $SE_1, \ldots, SE_4$ durch Infinite–Server–Stationen repräsentiert, deren mittlere Bedienzeiten den Antwortzeiten der Basisdienste entsprechen. Das Interesse des Planers liegt daher insbesondere in der Dimensionierung der technischen Aus-

prägung seiner Meta–Suchmaschine. Die Nachbearbeitung der von den Basisdiensten gelieferten Referenzen, d.h. das Download der Dokumente und das nachträgliche Ranking bzw. die Anpassung an die Suchoptionen der Meta–Suchmaschine, werden von dem Post–Processor P durchgeführt. Diese Nachbearbeitung der Referenzen ist jedoch nur unter den oben geschilderten Umständen notwendig. Im Modell wird dies dadurch berücksichtigt, daß von den Suchmaschinen $SE_1, \ldots, SE_4$ mit den festen Wahrscheinlichkeiten $p_1, \ldots, p_4$ zu dem Post–Processor P verzweigt wird. Ferner wird angenommen, daß jede der Basis–Suchmaschinen stets 20 Referenzen liefert. Der Ergebnisaufbereiter W übernimmt die abschließende Aufbereitung und Sortierung der Referenzen aller Basisdienste sowie die Übertragung des Suchergebnisses an den Benutzer. Aus Darstellungsgründen ist der Post–Processor im Modell mehrfach abgebildet. Da Suchanfragen nacheinander abgearbeitet werden müssen und aufgrund der meist recht hohen Last i.a. mehrere Server dazu benötigt werden, sind P und W FCFS–Stationen mit mehreren Bedienern. Auf die konkrete Anzahl der Bediener der Stationen P und W wird im folgenden Abschnitt 7.1.2 eingegangen.

7.1.2 Analyse

Zur Analyse des Modells schätzt der Planer der Meta–Suchmaschine zunächst die von ihm nicht beeinflußbaren Modellparameter ab. Dies sind die Antwortzeiten der Basis–Suchmaschinen. Der Planer geht von negativ–exponentiell verteilten Antwortzeiten der Basisdienste $SE_1, \ldots, SE_4$ mit den Mittelwerten $1/\mu_1 = 1/\mu_2 = 1.0$ und $1/\mu_3 = 1/\mu_4 = 1.2$ (Sekunden) aus. Angemerkt sei an dieser Stelle, daß die Anzahl täglicher Suchanfragen an typische Basis–Suchmaschinen etwa zwischen 10 Mio. und 100 Mio. liegt. Ein absolute Spitzenposition nimmt Google ein. Google wurde 1998 an der Stanford University von L. Page und S. Brin [25] entwickelt und bearbeitet mittlerweile 200 Mio. Suchanfragen pro Tag (Stand Aug. 2005). Dazu werden mehr als 10.000 Linux–Server eingesetzt
(Quelle *www.symweb.de/glossar/google__285.htm*).

Die Anzahl täglicher Anfragen an eine Meta–Suchmaschine liegt i.a. deutlich unter der Anzahl der Anfragen an eine Basis–Suchmaschine. Im Fall der von der Universität Hannover betriebenen deutschsprachigen Meta–Suchmaschine MetaGer liegt diese Zahl bei etwa 100.000 Anfragen pro Tag (vgl. www.metager.de). Mit dieser Anzahl wird auch obiges Modell belastet. Zudem wird angenommen, daß die Anzahl Poisson–verteilt ist. Die Wahrscheinlichkeiten p_i, mit denen eine Nachbearbeitung der Referenzen durch den Post–Processor notwendig ist, schätzt der Planer mit $p_1 = p_2 = 0.3$ und $p_3 = p_4 = 0.4$ ein.

Die Entscheidung, die der Planer/Betreiber der Meta–Suchmaschine zu treffen hat, liegt in der Auswahl geeigneter Server–Hardware hinsichtlich Geschwindigkeit und Anzahl. Im Falle des Ergebnisaufbereiters ist diese Entscheidung eher unkritisch, da die Ergebnisaufbereitung im Sortieren einer relativ geringen Anzahl von Referenzen je Anfrage besteht. Im Modell sind offensichtlich je Anfrage 80 Referenzen zu sortieren, da jede der Suchmaschinen, wie oben bereits erwähnt, etwa 20 Referenzen liefert. Der Planer rechnet je Anfrage mit einer Bearbeitungzeit von etwa 0.6 Sekunden (negativ–exponentiell verteilt). Es ist leicht einzusehen, daß für diese Aufgabe ein Server reicht. Um jedoch auch einer steigenden Belastung gerecht werden zu können, setzt der Meta–Suchmaschinen Betreiber aufgrund der eher geringen Hardwarekosten 2 Server ein. Kritischer wirkt sich die Hardwareentscheidung im Falle des

Post–Processors aus. Die Anzahl der Referenzen, die der Post–Processor P täglich bearbeiten muß, liegt bei

$$(p_1 + p_2 + p_3 + p_4) * 20 * 100.000 = 2.800.000.$$

Zur Bewältigung dieser Last ist grob skizziert eine Entscheidung für eine geringe Anzahl schneller und teurer Server oder für eine größere Anzahl langsamer und günstiger Server zu treffen. Der Betreiber der Meta–Suchmaschine hat die Möglichkeit, sich zwischen drei unterschiedlichen Hardware–Lösungen zu entscheiden. In den drei Fällen liegt die mittlere Bearbeitungszeit des Post–Processors je Referenz bei 0.8, 1.0 bzw. 1.2 Sekunden, wobei ebenfalls eine negativ–exponentiell verteile Zeitdauer angenommen wird. Die Bearbeitungsraten μ der Bedienzeit der Station P liegen somit bei 1.25, 1.0 bzw. 0.833. Aufgrund des Auslastungsgesetzes ergeben sich damit leicht Anforderungen hinsichtlich der minimalen Anzahl K_{min} notwendiger Server. Diese Anzahl beträgt:

$$K_{min} = \lceil (100.000 * 1.7 * 20)/\mu \rceil \tag{7.1}$$

Da jedoch andererseits die Anforderungen hinsichtlich der Antwortzeiten von (Meta–) Suchmaschinen und damit die Akzeptanz derselben sehr hoch sind, ist die Anzahl K der Server zu ermitteln, die zur Einhaltung gewisser durchschnittlicher Antwortzeiten nötig ist. Zur Analyse obigen Modells werden zunächst weitere Vereinfachungen vorgenommen. Aufgrund der Poisson–verteilten Anzahl täglicher Anfragen an die Meta–Suchmaschine ist die Anzahl an Bearbeitungsaufträgen für den Post–Processor ebenfalls Poisson–verteilt. Die durchschnittliche Anzahl beträgt $(p_1+p_2+p_3+p_4)*100.000 = 140.000$ und besteht jeweils aus 20 Referenzen. Im Modell wird vereinfachend angenommen, daß der Post–Processor statt der 140.000 Aufträge zu je 20 Referenzen 2.800.000 Einzelaufträge (ebenfalls Poisson–verteilt) erhält. Daraus läßt sich für die unterschiedlichen Bedienraten μ und unterschiedliche Server–Anzahlen K jeweils die Antwortzeit des Post–Processors ermitteln. Diese Antwortzeit wird anschließend genutzt, um die ersten beiden Momente der Durchlaufzeiten der vier isoliert voneinander betrachteten parallelen Stränge des Fork/Join–Netzes in Abbildung 7.3 zu ermitteln. Im nächsten Schritt werden die Momente der Durchlaufzeit zur Bestimmung von Aggregaten genutzt. An dieser Stelle sei angemerkt, daß in diesem Schritt lediglich zwei Aggregate bestimmt werden müssen, da je zwei der vier parallelen Stränge identisch sind. Schließlich werden die parallelen Teilnetze durch die Aggregate ersetzt, und das Modell ließe sich anhand des Analyseverfahrens für erweiterte Fork/Join–Warteschlangennetze analysieren. Dieser direkte Weg wurde jedoch aus Aufwandsgründen nicht beschritten. In Abschnitt 5.2.4 wurde bereits erwähnt, daß die Effizienz des Analyseverfahrens für das Upper–Bound Modell hinsichtlich Rechenzeit und Speicherplatzbedarf mit steigender Anzahl paralleler Bediener deutlich abnimmt. Die direkte Analyse des vorliegenden Modells mit vier parallelen Bedienern ergab Kardinalitäten der Makrozustandsräume $\tilde{Z}_k$ (vgl. Gl. (5.9) in Abschnitt 5) im Bereich zwischen 1372 und 2916 Zuständen. Zur Bestimmung der Größe des QBDs für das Upper–Bound Modell bzw. zur Bestimmung der Größe der involvierten Matrizen, ist die Kardinalität der Makrozustandsräume mit dem Produkt der Phasenanzahlen des Ankunftprozesses und der Bedienprozesse zu multiplizieren. Aufgrund der negativ–exponentiell verteilten Zwischenankunftszeiten besitzt der Ankunftsprozeß lediglich eine Phase. Die Phasenanzahlen der Bedienzeitverteilungen ergeben sich aus den berechneten Aggregaten. Insgesamt ergaben sich Zustandsraumgrößen des Modells zwischen 22.000 und 180.000 Zuständen. Die Verwendung der matrix–geometrischen Methoden

auf derart großen Zustandsräumen ist unter Effizienzaspekten nicht mehr sinnvoll. Daher wurde zur Gewinnung der nachfolgend dargestellten Analyseresultate die in Abschnitt 6.4 erwähnte Divide–and–Conquer Methode angewandt. Dazu wurde zunächst in einem ersten Schritt aus jeweils zwei der parallelen Stränge ein Fork/Join–Modell mit zwei parallelen Bedienern gebildet. Da die beiden ersten und die beiden letzten Stränge identisch sind, reicht bei Betrachtung des ersten und dritten bzw. zweiten und vierten Strangs in diesem ersten Schritt ein einziges Fork/Join–Modell aus. Für dieses Modell werden die ersten beiden Momente der Durchlaufzeitverteilung ermittelt. Auf dieser Grundlage wird ein Aggregate gebildet. Aus diesem Aggregat wird im nächsten Schritt wiederum ein Fork/Join–Modell mit zwei homogenen parallelen Bedienern gebildet. Auf der Grundlage dieses Fork/Join–Modells läßt sich schließlich das Dekompositionsverfahren zur Analyse des Modells aus Abbildung 7.3 einsetzen.

Auf diese Weise konnte der Analyseaufwand gegenüber dem Fork/Join–Modell mit vier parallelen Bedienern erheblich reduziert werden. Die Zustandsraumgrößen der QBDs für die Fork/Join–Modelle im ersten Schritt lagen zur Gewinnung der nachfolgend beschriebenen Analyseresultate im Bereich zwischen 52 und 160 Zuständen. Die Zustandsraumgrößen der QBDs für die im zweiten Schritt betrachteten Fork/Join–Modelle lagen zwischen 225 und 325 Zuständen. Derartige Zustandsräume lassen sich anhand der matrix-geometrischen Methoden leicht behandeln. Die im folgenden dargestellten Analyseresultate wurden auf einem AMD Opteron Prozessor mit 2 GHz, 1 MB Cache und 6 GB Hauptspeicher ermittelt und lagen jeweils nach wenigen Minuten vor. Auf die Approximationsfehler, die durch dieses Vorgehen induziert werden, wird am Ende dieses Abschnittes eingegangen.

Die Abbildung 7.4 stellt die mittleren Antwortzeiten der Meta–Suchmaschine in Abhängigkeit der drei unterschiedlichen Hardware–Lösungen und unterschiedlichen Serveranzahlen K dar. Anhand der Abbildung läßt sich leicht die minimal erreichbare Antwortzeit der Meta–Suchma-

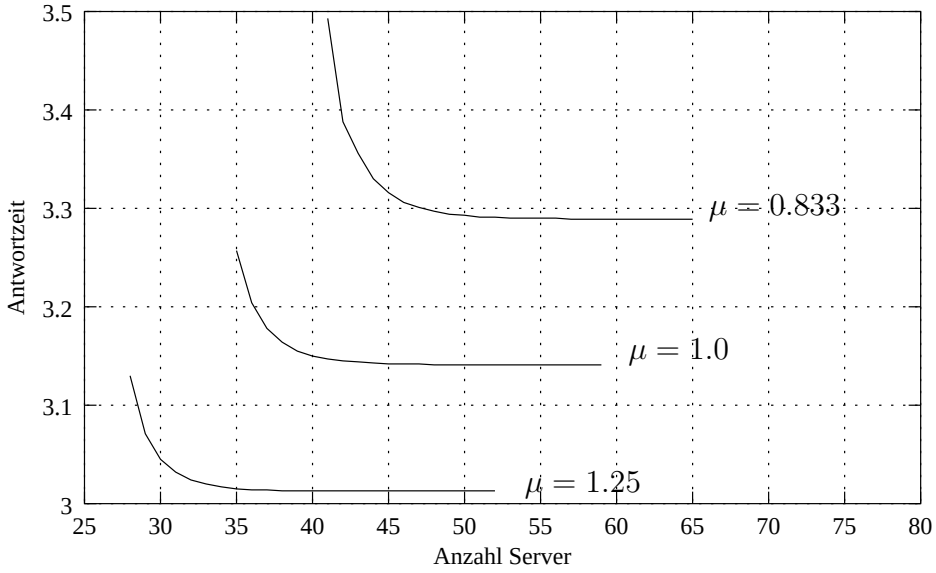

Abbildung 7.4: Antwortzeit der Meta–Suchmaschine in den Fällen $\mu = 1.25, 1.0, 0.833$

schine in den drei Hardwarekonfigurationen bei ausreichender Serveranzahl ablesen. Sicherlich ist in allen drei Fällen die erzielbare Antwortzeit akzeptabel. Die Resultate können somit als

Grundlage zur Berechnung der Kosten für die drei unterschiedlichen Lösungen hernagezogen werden.

In die Entscheidung für eine der Konfigurationen wird jedoch ein zusätzliches Argument mit aufgenommen. Dazu wird angenommen, daß der Nutzungsgrad einer Internet–Suchmaschine halbjährlich um etwa 10 % zunimmt. Dazu wurden weitere Untersuchungen an obigem Modell durchgeführt, bei denen die Belastung durch Suchanfragen um 10, 21, 33, 46 und 61 % erhöht wurde. Dabei wurde davon ausgegangen, daß auch die Betreiber der Basisdienste $SE_1, \ldots, SE_4$ bestrebt sind, auf steigende Belastungen ihrer Suchmaschinen geeignet zu reagieren, so daß deren Antwortzeiten unverändert bleiben. Die Abbildungen 7.5, 7.6 7.7 stellen die Ergebnisse dieser Analysereihen für die Fälle $\mu = 1.25$, $\mu = 1.0$ und $\mu = 0.833$ dar. Aufgetragen sind jeweils die Resultate für die erwartete Ausgangslast von 100.000 Suchanfragen pro Tag sowie die Resultate bei 10, 21, 33, 46 und 61 % erhöhter Last.

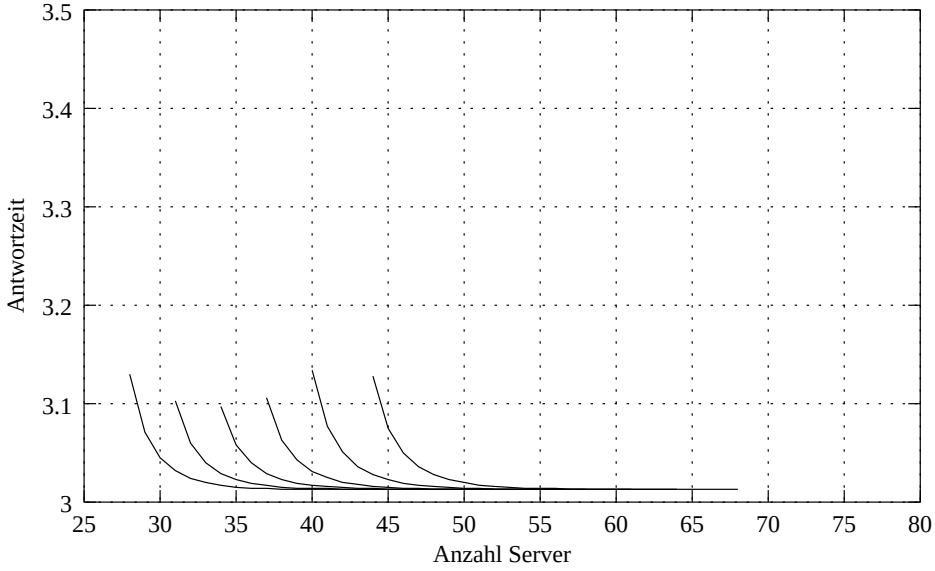

Abbildung 7.5: Antwortzeit der Meta–Suchmaschine bei steigender Last, $\mu = 1.25$

Angemerkt sei an dieser Stelle, daß sämtliche Analysereihen mit einer minimalen Anzahl an Servern von $K_{min}+2$ durchgeführt wurden. Daher kann der Fall eintreten, daß bei konstantem μ und der Wahl $K = K_{min}+2$ die Station P im Fall einer höheren Ankunftsrate eine niedrigere Auslastung besitzt, als im Fall einer geringeren Ankunftsrate. Dies wirkt sich unmittelbar auf die Antwortzeit der Station P und damit auf die Antwortzeit der Meta–Suchmaschine aus. In den Abbildungen äußert sich diese Tatsache in den nicht notwendigerweise steigenden linken Anfangswerten der Kurven. Ferner wurden zur besseren Vergleichbarkeit der Resultate in allen Fällen identische Intervalle der Koordinaten–Achsen gewählt.

Abschließend bleibt die Approximationsgüte der Analyseresultate zu beurteilen. Dazu wurden sämtliche Analyseresultate mit je einer Simulation des entsprechenden Modells verglichen. Die Simulationsläufe wurden jeweils nach Erreichen eines 95%–Konfidenzintervalls der Breite 5% für die Antwortzeit der Meta–Suchmaschine abgebrochen. Da in allen Fällen im wesentlichen die gleichen Approximationsfehler auftraten, werden im folgenden lediglich die Resultate für den Fall der Bedienrate $\mu = 0.8$ der Station P und der ursprünglichen Last von 100.000

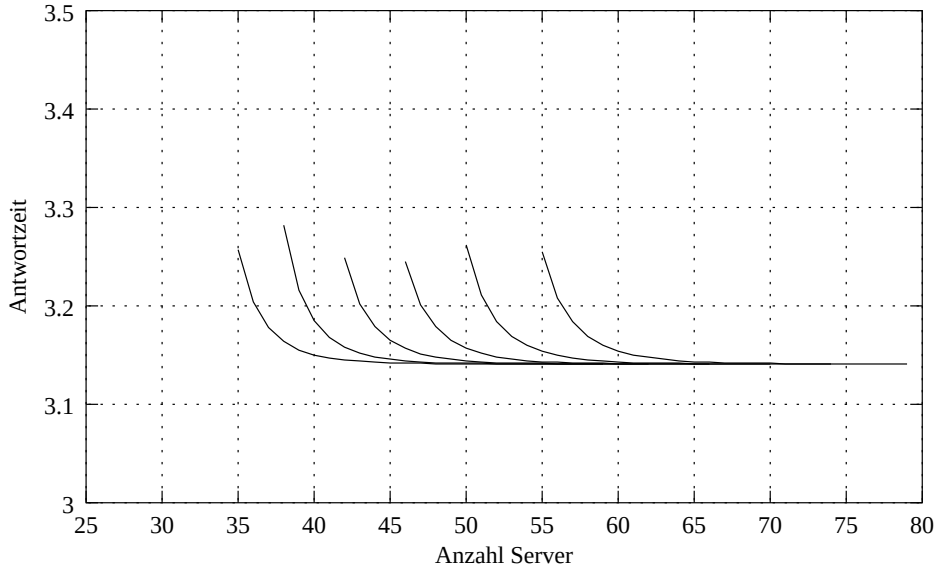

Abbildung 7.6: Antwortzeit der Meta–Suchmaschine bei steigender Last, $\mu = 1.0$

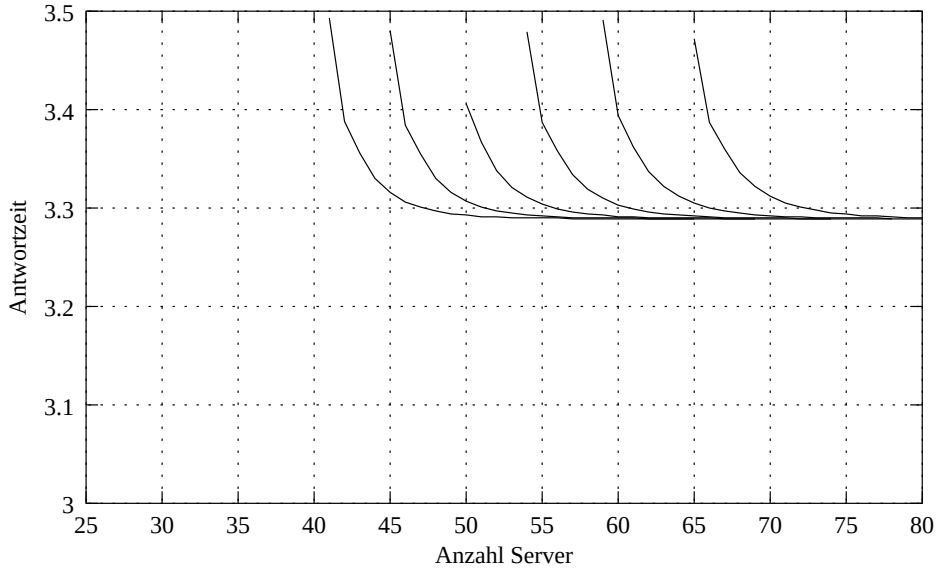

Abbildung 7.7: Antwortzeit der Meta–Suchmaschine bei steigender Last, $\mu = 0.833$

Suchanfragen pro Tag betrachtet. Die Tabelle 7.1 stellt für diesen Fall die im Vergleich zu der Simulation auftretenden Approximationsfehler dar. Aufgetragen sind jeweils in Abhängigkeit der Anzahl der Server die Resultate des in dieser Arbeit vorgestellten Analyseverfahrens (UB), die Resultate einer Simulation sowie die prozentuale Abweichung beider Analyseresultate.

Anzahl Server	UB	Sim	Δ (%)
28	3.13	3.56	12.19
29	3.07	3.46	11.33
30	3.04	3.42	11.03
31	3.03	3.40	10.94
32	3.02	3.39	10.90
33	3.02	3.38	10.59
34	3.02	3.37	10.50
35	3.02	3.37	10.50
36	3.01	3.37	10.49
37	3.01	3.36	10.40
38	3.01	3.36	10.40

Tabelle 7.1: Approximationsgüte der Analyseresultate

Die Ursache für diese Approximationsfehler liegt in der Anwendung der Divide–and–Conquer Methode zur Analyse des Fork/Join–Teilmodells. Der deutlich reduzierte Analyseaufwand hat somit einen erhöhten Approximationsfehler zur Folge. Es sei jedoch angemerkt, daß der auftretende Approximationsfehler von etwa 10% in einem akzeptablen Rahmen liegt. Somit kann der Planer der Meta–Suchmaschine obige Analyseresultate in seine Entscheidung zur Wahl der Hardwareausstattung mit einbeziehen. So läßt sich z.B. ablesen, bis zu welcher Belastung die konkrete Wahl einer bestimmten Server–Anzahl mit einer der genannten Geschwindigkeiten akzeptable Antwortzeiten liefert.

Dieses Beispiel demonstriert deutlich die Relevanz der in dieser Arbeit vorgestellten Analysemethode für erweiterte Fork/Join–Warteschlangennetze. Insbesondere die hohen Anforderungen hinsichtlich der Antwortzeit an zahlreiche, teilweise oben genannte Web–Dienste, die in ihrem prinzipiellen Aufbau der hier dargestellten Meta–Suchmaschine ähnlich sind, machen eine geeignete Modellierung und Analyse notwendig. Dazu ist das um die Analyse von Fork/Join–Netzen angereicherte Dekompositionsverfahren zur Grobanalyse derartiger Modelle gut geeignet. Im folgenden Kapitel wird die Relevanz des Verfahrens im Kontext von Logistiknetzen demonstriert.

Kapitel 8

Logistik

Dieses Kapitel stellt das Anwendungsfeld Logistik als weiteres interessantes und hochaktuelles Beispiel für den Einsatz der in der vorliegenden Arbeit entwickelten Analysemethode für Fork/Join–Warteschlangennetze vor. Die Logistik nimmt in der modernen Industriegesellschaft einen zentralen Stellenwert ein. Die Aufgaben der Logistik liegen in der Planung, Steuerung, Optimierung und Kontrolle des gesamten Material– und Informationsflusses entlang der Wertschöpfungsketten eines Unternehmens. Die Komplexität dieser Aufgabe wird anhand gewisser Rahmenbedingungen deutlich, die diesen Aufgaben zugrundeliegen. Insbesondere die Verschärfung des Wettbewerbs aufgrund der Globalisierung, sich ständig ändernden Marktvorgaben und der Verfügbarkeit neuer Technologien stellt eine enorme Herausforderung dar. Wertschöpfungsprozesse müssen folglich flexibel ausgerichtet sein, um schnell auf schwankende Nachfragesituationen, individuelle Kundenwünsche, kürzere Lieferzeiten und striktere Lieferterminzusagen sowie auf höhere Anforderungen hinsichtlich der Produkt– und Servicequalität reagieren zu können. Dabei müssen stets Unternehmens–interne Ziele wie optimale Kapazitätsauslastungen, minimale Kosten, niedrige Bestände und kurze Durchlaufzeiten berücksichtigt werden. Diese Umstände führten in der Vergangenheit zu einem strukturellen Veränderungsprozeß der Industrielandschaft. Um den Anforderungen der schnellebigen Märkte gerecht zu werden und flexibel auf veränderte Rahmenbedingungen reagieren zu können, um innovativ zu sein und schnell neue Produkte entwickeln und vermarkten zu können, um schließlich Wettbewerbsvorteile zu erzielen, müssen sich Unternehmen auf ihre Kernkompetenzen konzentrieren und weniger effiziente Geschäftsfelder und Teilbereiche an spezialisierte Partner outsourcen. Dies führt letztlich zu einer Verringerung der Wertschöpfungstiefe. Nach Koether [51] kauften Industriebetriebe im Jahr 2004 etwa 50% ihres Verkaufswertes hinzu mit steigendem Trend. Andererseits bedeutet diese Entwicklung gleichzeitig, daß der Erfolg eines Unternehmens nicht ausschließlich in der eigenen Verantwortung liegt, sondern zudem von der sorgfältigen Auswahl der Partner und deren Erfolg abhängt, kurzum von einem optimalen Funktionieren der gesamten Kollaboration.

Aus dem Blickwinkel der Logistik hat dies zur Konsequenz, daß es nicht ausreicht, die Wertschöpfungsprozesse eines Unternehmens isoliert zu betrachten. Eine Wertschöpfungskette endet nicht an Unternehmensgrenzen, sondern erstreckt sich von der Auftragserteilung eines Kunden über die verschiedenen Stufen der Kollaboration bis hin zur Auslieferung des

fertigen Produktes an den Kunden. Zur Unterstützung der Aufgaben der Logistik ist somit ein Instrumentarium notwendig, das die gesamte Logistikkette von der Auftragserteilung seitens des Kunden über sämtliche Zulieferprozesse, dem Produktionsprozeß bis hin zur Auslieferung des fertigen Produkts an den Kunden abbildet. Dieses Instrumentarium muß gleichzeitig geeignet sein, Schwachstellen in der Logistikkette aufzudecken und Aussagen hinsichtlich der Kernziele der Logistik, d.h. Kapazitätsauslastung, Durchlaufzeiten, Bestände und Servicegrade abzuleiten. Zu diesem Zweck haben sich Prozeßketten etabliert, auf die im folgenden Abschnitt genauer eingegangen wird.

8.1 Prozeßketten in der Logistik und das ProC/B–Toolset

Die integrierte, werkzeuggestützte Planung, Implementierung, Steuerung und Kontrolle von Logistiknetzen und Geschäftsprozessen erfordert eine einheitliche und korrekt formalisierte Beschreibungssprache. In der Praxis haben zu diesem Zweck Prozeßketten ein hohes Maß an Akzeptanz gefunden. Prozeßketten beschreiben betriebliche und unternehmensübergreifende Abläufe in zeitlich-logischer Reihenfolge und berücksichtigen sämtlich Informations– und Materialflüsse.

Am Institut für Wirtschaftinformatik der Universität des Saarlandes wurde im Jahr 1991 die „Ereignisgesteuerte Prozeßkette (EPK)“ [46] entwickelt und im Jahr 1993 von der IDS Prof. Scheer GmbH (heute AG) in dem ARIS–Toolset implementiert [80, 78]. ARIS unterstützt Unternehmen bei der Modellierung, Analyse und Optimierung von Prozessen. Die Fachgruppe „Geschäftsprozeßmanagement mit Ereignisgesteuerten Prozeßketten“ der Gesellschaft für Informatik eV (GI) beschäftigt sich zudem mit der Formalisierung der Syntax und Semantik und der Analyse von EPKs [85, 48].

Die Grundlage für die in den nachfolgenden Abschnitten vorgestellten Anwendungsfälle bildet das Prozeßketten–Instrumentarium nach Kuhn [58], entwickelt und erfolgreich eingesetzt am Fraunhofer Institut für Materialfluß und Logistik (IML) in Dortmund. Dieses Instrumentarium findet gleichzeitig Anwendung in dem Sonderforschungsbereich 559 „Modellierung großer Netze in der Logistik“ (SFB 559), der im Jahr 1998 an der Universität Dortmund in Kooperation mit dem Fraunhofer IML eingerichtet wurde (*www.sfb559.uni-dortmund.de)*. Der SFB 559 verfolgt das Ziel, eine Theorie zur Gestaltung, Organisation und Steuerung großer Logistiknetze zu entwickeln. In diesem Kontext bildet das Prozeßketten–Instrumentarium eine gemeinsame Sprache, auf deren Basis die beteiligten interdisziplinären Fachrichtungen aus Maschinenbau, Logistik, Betriebswirtschaft, Statistik und Informatik miteinander kommunizieren. Innerhalb dieses Projekts befaßt sich der Lehrstuhl Informatik IV für „Modellierung und Simulation“ u.a. mit der Konkretisierung und Formalisierung der Prozeßketten nach Kuhn derart, daß sie einer automatisierten Analyse zugänglich sind. Dazu wird ausgenutzt, daß Prozeßketten i.w. diskrete ereignisorientierte dynamische Systeme (DEDS) modellieren, für die in der Informatik ein reichhaltiges Angebot unterschiedlicher Analyseverfahren bereitsteht. Auf dieser Grundlage ist am Lehrstuhl Informatik IV das *ProC/B*–Toolset [6, 3, 7] entstanden, das die integrierte Modellierung und Analyse von Prozeßketten ermöglicht.

Im folgenden werden die Prozeßketten, wie sie im Pro/C–Toolset formalisiert und implemen-

tiert sind[1], erläutert. Zudem wird das ProC/B–Toolset mit seiner Funktionalität selbst kurz vorgestellt. Das Ziel der Ausführungen liegt in der Vermittlung eines grundlegenden Verständnisses für die nachfolgenden Beispiele. Daher wird auf eine detaillierte Darstellung verzichtet und auf [6, 3] verwiesen.

Prozeßketten stellen betriebliche und unternehmensübergreifende Prozesse in ablauflogischer Reihenfolge dar. Sie beginnen stets mit einer Quelle, die die zu bewältigende Auftragslast beschreibt und enden mit einer Senke, die die Erfüllung eines Auftrags bzw. das Prozeßende repräsentiert. Zwischen der Quelle und der Senke beschreiben i.a. mehrere Prozeßkettenelemente (PKE) eine Abfolge von Aktivitäten, die zur Prozeßabwicklung nötig sind. Die ablauflogische Reihenfolge der PKEs wird durch Konnektoren spezifiziert, die die parallele, die alternative oder die sequentielle Ausführung von PKEs bzw. Teilprozessen (Folge von PKEs) erzwingen. Die PKEs enthalten neben der reinen Beschreibung einer Aktivität zusätzliche Informationen darüber, welche Kapazitäten/Ressourcen in welcher Anzahl und ggfls. wie lange für die Aktivität benötigt werden. Aktivitäten können zeitverbrauchend bzw. platzeinnehmend sein, indem sie aktive Ressourcen (Personal, Maschinen, etc.) für eine gewisse Zeit in Anspruch nehmen bzw. eine bestimmte Menge an passiven Ressourcen (Lagerplatz) belegen. Alternativ zur Nutzung elementarer aktiver bzw. passiver Ressourcen können PKEs durch sog. Funktionseinheiten (FE) detaillierter spezifiziert werden. FEs beschreiben komplexe Abläufe ebenfalls in Form der zuvor beschriebenen Prozeßketten. In diesem Sinne erhalten Prozeßketten eine hierarchische Struktur, bei der FEs auf unterer Hierarchiestufe Diensterbringer für FEs auf der darüberliegenden Hierarchiestufe sind.

Das ProC/B–Toolset unterstützt die integrierte Modellierung und Analyse von Prozeßketten. In dem ProC/B–Editor lassen sich sog. ProC/B–Modelle erstellen, die die dynamische und statische Sicht eines Modells unterscheiden. Die dynamische Sicht besteht aus einer oder mehreren der soeben beschriebenen Prozeßketten. Die statische Sicht spezifiziert die Anzahl und Ausprägung der in dem Modell verfügbaren Ressourcen. Insbesondere enthält sie die Definition von Funktionseinheiten, so daß die statische Sicht die Modellhierarchie widerspiegelt. Die Abbildung 8.1 zeigt exemplarisch ein ProC/B–Modell.

Im Zentrum der Darstellung steht die Beschreibung einer Prozeßkette mit Quelle, Senke und sechs PKEs. Der untere Teil der Modellbeschreibung enthält die Spezifikation der verfügbaren Ressourcen bzgl. ihrer Typen, ihrer Anzahlen und ihrer genauen Ausprägung. In der Quelle wird in gewissen zeitlichen Abständen jeweils ein Auftrag erzeugt. Die zeitlichen Abstände unterliegen einer Negativ–Exponentialverteilung mit der Rate „ein Auftrag pro Zeiteinheit (ZE)“. In einem ersten Schritt (PKE_1) wird der Auftrag von einem Mitarbeiter/Personal bearbeitet. Die Bearbeitungsdauer unterliegt im vorliegenden Fall ebenfalls einer Negativ–Exponentialverteilung mit der Rate $2/ZE$. Anschließend (PKE_2) erfordert der Auftrag die Einlagerung gewisser Teile in ein Lager und benötigt dafür fünf Lagerplätze. In dem (PKE_3) werden Aktivitäten durchgeführt, deren genaue Gestalt in einer unterliegenden FE *SubFE* spezifiziert ist. Im nächsten Schritt (PKE_4) werden fünf Teile aus dem Lager entnommen. Die senkrechten Konnektoren (UND–Konnektoren) haben die Bedeutung, daß die eingeschlossenen PKEs PKE_5 und PKE_6 gleichzeitig beginnen und beide Aktivitäten beendet sein müssen, bevor nachfolgende Aktivitäten beginnen können. Die UND–Konnektoren entsprechen in ihrer Semantik somit den in dieser Arbeit behandelten Fork/Join–Stationen. Der Typ

[1] Mit dem Begriff *Prozeßkette* sind nachfolgend stets diejenigen in der Formalisierung nach ProC/B gemeint.

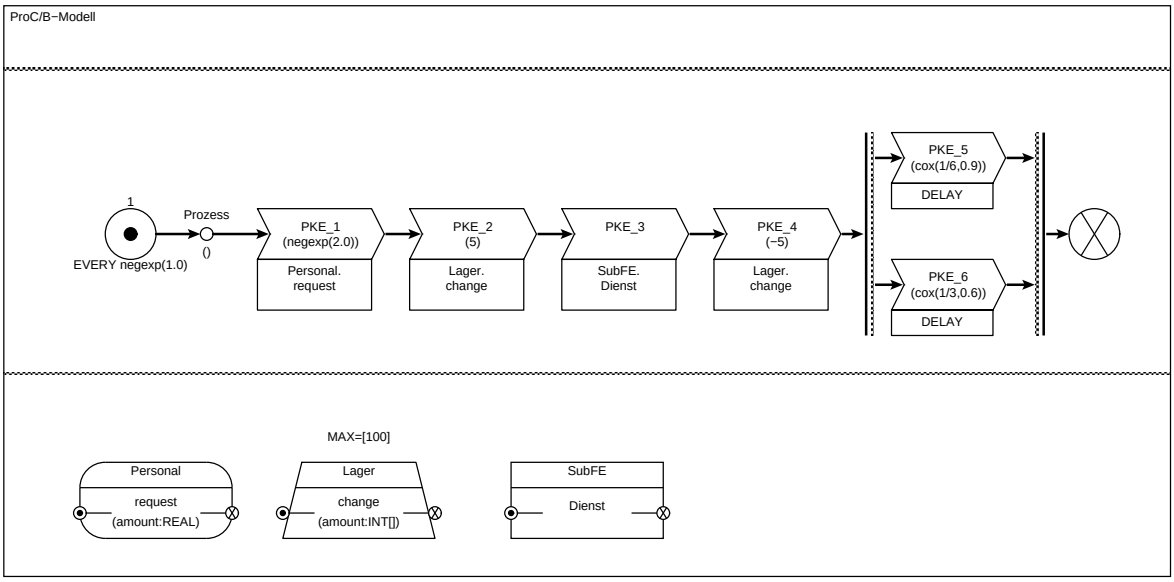

Abbildung 8.1: Beispiel eines ProC/B–Modells

Delay der beiden PKEs zeigt an, daß PKE_5 und PKE_6 nicht an Ressourcen gebunden sind, sondern lediglich einen zeitverzögernden Einfluß haben. Die Zeitverzögerung ist durch eine Cox–Verteilung anhand der Rate und des Variationskoeffizienten spezifiziert. Die dynamische und statische Struktur von SubFE erfolgt auf dieselbe Weise, wie die Beschreibung eines ProC/B–Modells durch Angabe der Prozeßketten und der verfügbaren Ressourcen.

Derartige Prozeßketten beschreiben *Diskrete Ereignisorientierte Dynamische Systeme (DEDS)* [27]. Somit sind sie dem reichhaltigen Vorrat an effizienten qualitativen und quantitativen Analyseverfahren in diesem Bereich zugänglich. In dem ProC/B–Toolset sind sowohl zustandsraumbasierte numerische Verfahren als auch algebraische Warteschlangentechniken und Simulation integriert. Die Verfügbarkeit wird jeweils durch Übersetzer in entsprechend vorhandene Werkzeuge realisiert. Das Modellierungs– und Analysewerkzeug HIT [21, 22] wurde am Lehrstuhl Informatik IV „Modellierung und Simulation“ der Universität Dortmund entwickelt. Es enthält einen leistungsfähigen Simulator sowie spezielle zeit– und platzeffiziente Analyseverfahren für separable Warteschlangennetze. Durch einen Übersetzer, der ProC/B–Modelle in die Eingabesprache von HIT transformiert, werden sowohl der Simulator als auch die algebraischen Warteschlangentechniken im ProC/B–Kontext verfügbar. Die APNN–Toolbox [15, 16] wurde ebenfalls am Lehrstuhl IV entwickelt. Sie erlaubt die Modellierung und Analyse von Petri–Netzen und enthält ein reichhaltiges Repertoire an zustandsraumbasierten numerischen Analysetechniken. Ein entsprechender Modellübersetzer transformiert ProC/B–Modelle in Petri–Netze und macht numerische Analysetechniken damit in der ProC/B–Welt anwendbar. Schließlich ist auch das in dieser Arbeit entwickelte Verfahren durch einen Transformator an das ProC/B–Toolset angebunden. Dazu wurde im Rahmen dieser Arbeit ein Prototyp entwickelt, der die Analyse von erweiterten Fork/Join–Warteschlangennetzen anhand des Dekompositionsverfahrens nach Kühn erlaubt. Dieser Prototyp beinhaltet insbesondere die Analysemethode für Fork/Join–Warteschlangennetze.

8.2 Güterverkehrszentrum

Ein Güterverkehrszentrum (GVZ) ist ein logistischer Knotenpunkt, in dem Gütertransporte zwischen unterschiedlichen Verkehrsträgern umgeschlagen und zwischengelagert werden. Die Deutsche GVZ–Gesellschaft (*www.gvz-org.de*) charakterisiert ein GVZ durch

- die Ansiedlung verkehrswirtschaftlicher Betriebe, logistischer Dienstleister und logistikintensiver Industrie– und Handelsunternehmen in einem Gewerbegebiet.
- die Anbindung an mindestens zwei Verkehrsträger, insbesondere Straße/Schiene.
- Managementfunktionen lokaler GVZ–Gesellschaften, die ebenfalls kooperative Aktivitäten initiieren und moderieren.

Die Standortfrage ist somit ein wesentliches Kriterium für die Errichtung eines GVZ. Die Abbildung 8.2 zeigt eine Luftaufnahme des GVZ Dresden mit unmittelbarer Anbindung an den Schienengüterverkehr, an die Elbe–Container Linie und eine vierspurige Verbindung zu den Autobahnen A4/A17. Aus organisatorischer Sicht besteht ein GVZ im wesentlichen aus ei-

Abbildung 8.2: Luftaufnahme des GVZ Dresden (Quelle: www.gvz-dresden.de)

nem Terminal für den Kombinierten Ladungsverkehr (KV–Terminal) und einer Stückgutumschlaghalle (SUH), in denen die zentralen Operationen des Umschlags, der Lagerung, der

Auseinzelung von Gütern, der Komissionierung und schließlich wiederum der Auslieferung stattfinden.

Das Interesse dieses Abschnittes besteht insbesondere in der Kapazitätsplanung des KV–Terminals. Aufgrund eines steigenden Güteraufkommens insbesondere durch den Schienengüterverkehr waren zahlreiche deutsche GVZ–Betreiber in den letzten Jahren gezwungen, ihre Kapazitäten zu erhöhen. So verdoppelte die Betreibergesellschaft des KV–Terminals Lübeck–Travemünde im Jahr 2003 die Umschlagskapazität von/auf die Schiene auf jährlich 140.000 Trailer, Container oder Wechselbrücken. Im Jahr 2005 beginnen die Umbauarbeiten des KV–Terminals Leipzig zur Erhöhung der Kapazitäten von derzeit 55.000 auf 120.000 Container. Im KV–Terminal Rostock ist ein Ausbau der Kapazitäten von derzeit etwa 56.000 Einheiten auf 80.000 Einheiten geplant. Ausgelöst wird dieser Trend durch den überlasteten Straßenverkehr, die LKW–Maut, Bemühungen zum Umweltschutz sowie Regierungsanstrengungen zur Verlagerung des Straßengütertransports auf die Schiene. In diesem Abschnitt wird verdeutlicht, auf welche Weise die in dieser Arbeit entwickelte Analysemethode in der Planungsphase eines Neu– oder Umbaus eines KV–Terminals genutzt werden kann, die erreichbaren Kapazitäten zu beurteilen. Dabei wird es weniger auf eine detaillierte Abbildung des KV–Terminals in einem Modell ankommen, sondern vielmehr auf eine Grobeinschätzung anhand eines abstrakten Modells. In [2] wird ein Modell eines KV–Terminals vorgestellt, in dem der Umschlag zwischen Zügen und LKWs unter Einbeziehung eines Lagers mit fester Kapazität simulativ untersucht wird. In dem vorliegenden Ansatz besteht das Interesse an der Bewertung von technischen Maßnahmen zur Erhöhung der Kapazität eines KV–Terminals. Dazu wird im folgenden zunächst das angenommene Szenario beschrieben und im Anschluß die Analyse präsentiert.

8.2.1 Modellbeschreibung

Betrachtet wird der Container–Umschlagprozeß von Zügen im KV–Terminal. Container, die vom Zug abgeladen werden, werden in ein Lager verbracht. Umgekehrt wird der Zug mit Containern aus dem Lager beladen. Dieser Umschlagprozeß wurde häufig von mobilen Umschlaggeräten durchgeführt, die nacheinander die Container vom Zug abluden, um sie dann auf einer freien Stelle oder auf einem anderen Container zwischenzulagern. Nach Abschluß des Umschlagprozesses wurden die Container dann an ihren endgültigen Platz im Lager verbracht. Das Be– und Entladen eines Zuges dauerte auf diese Weise meist mehrere Stunden. Zudem bestand ein recht hoher Flächenbedarf zur Zwischenlagerung der Container. In modernen KV–Terminals werden schienengebundene Portalkräne eingesetzt, die entlang der Gleise fahren und den Umschlag von Containern zwischen den Zügen und einer Vorsortierfläche realisieren. Parallel dazu transportieren Lagerkräne Container zwischen Vorsortierfläche und Hochregallager. Dieses Szenario ist in der Abbildung 8.3 skizziert.

Das Interesse dieses Abschnitts liegt in der Ermittlung der Last, die auf diese Weise von dem KV–Terminal zu bewältigen ist. Zur Beantwortung dieser Fragestellung wird das in der Abbildung 8.4 dargestellte ProC/B–Modell betrachtet. In dem Modell wird angenommen, daß für die Umschlagprozesse zwei Portalkräne und zwei Lagerkräne zur Verfügung stehen und daß jeder Zug 30 Container transportiert. Nach der Ankunft eines Zuges setzen beide Portalkräne je einen Container vom Zug auf die Vorsortierfläche um. Parallel dazu holen beide Lagerkräne

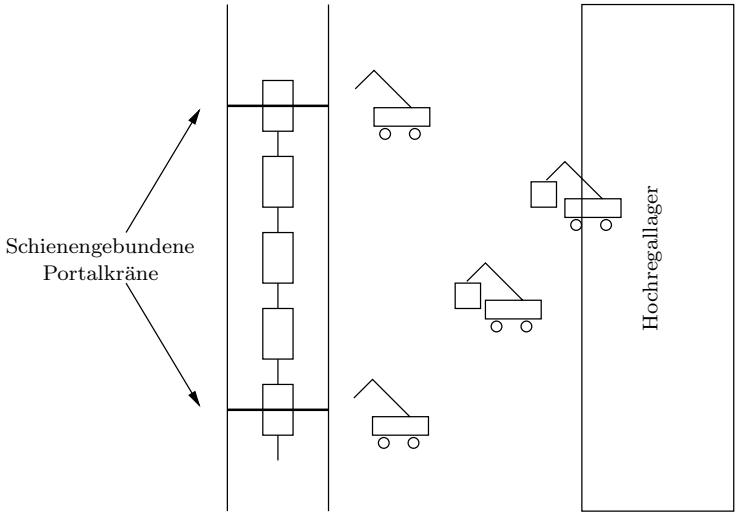

Abbildung 8.3: Skizze des Umschlagterminals

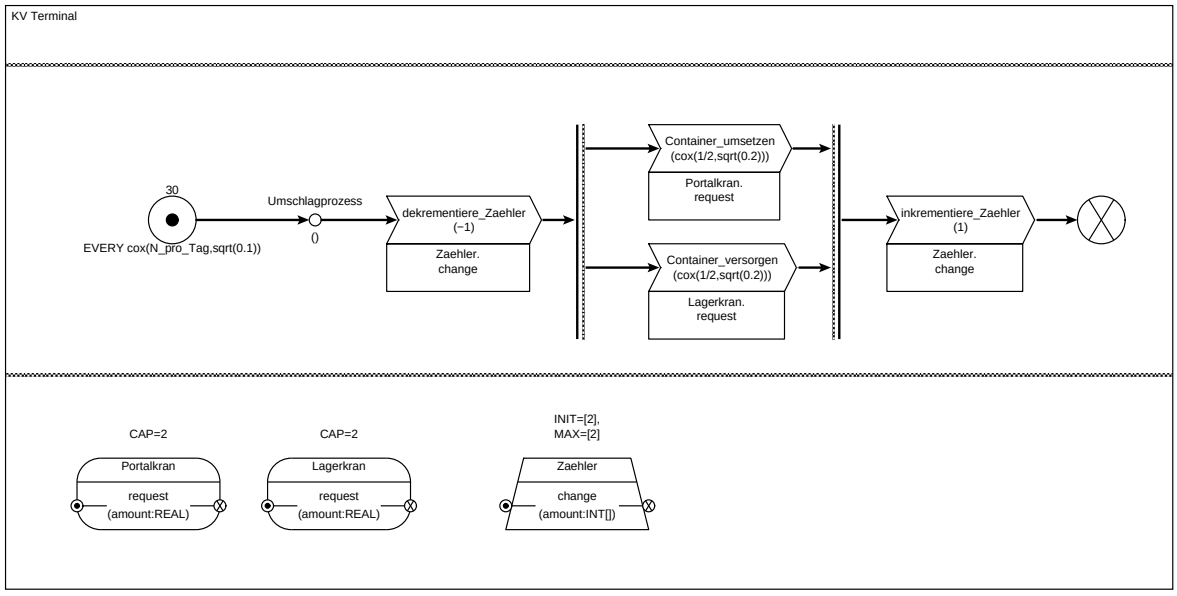

Abbildung 8.4: ProC/B–Modell des KV–Terminals

je einen Container aus dem Hochregallager und stellen diesen auf der Vorsortierfläche ab. Anschließend verbringen die Portalkräne die von den Lagerkränen bereitgestellten Container auf den Zug. Gleichzeitig verbringen die Lagerkräne die von den Portalkränen bereitgestellten Container in das Hochregallager. Auf dieselbe Weise setzen die Portalkräne und Lagerkräne ihre Tätigkeit mit den übrigen Containern fort.

Zu diesen Ausführungen bzw. zu dem in der Abbildung 8.4 dargestellten Modell sind zwei Anmerkungen hinzuzufügen. Die erste Anmerkung betrifft die Prozeßkettenelemente *dekrementiere_Zaehler* und *inkrementiere_Zaehler*. Da in dem Modell von zwei Portalkränen und zwei Lagerkränen ausgegangen wird, stellen diese Prozeßkettenelemente sicher, daß stets höchstens zwei Umschlagvorgänge gleichzeitig stattfinden. Die zweite Anmerkung bezieht sich auf die Tatsache, daß das Modell der Abbildung 8.4 streng genommen nicht exakt mit obiger Beschreibung übereinstimmt. Der Synchronisationspunkt in dem Modell, und damit das Ende eines Umschlagvorgangs liegt an der Vorsortierfläche. Dies liegt daran, daß an dieser Stelle die Portalkräne bzw. die Lagerkräne darauf warten, daß der jeweilige Gegenpart einen Container aus dem Hochregallager bzw. vom Zug geholt und diesen auf der Vorsortierfläche abgestellt hat. An dieser Stelle beginnt folglich auch jeder Umschlagvorgang, was im Fall des jeweils ersten Containers eines Zug gemäß obiger Beschreibung nicht korrekt ist. Ebenso endet der Umschlagvorgang des jeweils letzten Containers eines Zuges nicht auf der Vorsortierfläche, so daß auch in diesem Fall das Modell streng genommen nicht korrekt ist. Hinsichtlich der Analyse des Modells haben diese geringen Unstimmigkeiten jedoch keinen Einfluß.

Schließlich wird in dem Modell angenommen, daß die Zeitdauer, die ein Portalkran für einen Umschlagvorgang benötigt, etwa zwei Minuten beträgt und einer Schwankung unterliegt, die durch den Variationskoeffizienten $\sqrt{0.2}$ erfaßt wird. Die gleiche Annahme wird auch im Fall eines Umschlagvorgangs der Lagerkräne getroffen. Im folgenden Abschnitt wird auf die Analyse dieses Modells eingegangen.

8.2.2 Analyse

Wie bereits ausgeführt, besteht das Ziel der Untersuchungen an obigem Modell darin, die von dem System täglich bewältigbare Last zu ermitteln. Dazu wird im folgenden ein minimaler Ankunftstakt bzw. äquivalent dazu eine maximale Ankunftsrate der Züge bestimmt. Dabei wird die Nebenbedingung beachtet, daß Züge kostenintensive Verkehrsmittel sind und daher geringe Wartezeiten haben sollten. Insbesondere sollte der Umschlagprozeß unmittelbar nach der Ankunft eines Zuges beginnen und nicht durch den laufenden Umschlagprozeß eines vorangegangenen Zuges verzögert werden. Alternativ formuliert sollte sich durchschnittlich höchstens ein Zug in dem KV–Terminal befinden.

Zur Analyse dieses Modell anhand der Fork/Join–Warteschlangennetze wird zunächst eine Vereinfachung vorgenommen, die die betrachteten Leistungsobjekte betrifft. In obigem Modell werden mit der Ankunft eines Zuges unmittelbar 30 Container in das System eingebracht, die dann einzeln umgeschlagen werden. In Warteschlangennetzen werden derartige Situationen mit dem Begriff *Batch–Ankünfte* beschrieben bzw. modelliert. An dieser Stelle sei angemerkt, daß sich das Upper–Bound Modell prinzipiell um diese Eigenschaft erweitern ließe, aus praktischer Sicht führte diese Erweiterung jedoch zu einer deutlichen Erhöhung

des Analyseaufwands. In der vorliegenden Arbeit werden Batch–Ankünfte nicht berücksichtigt, und daher wird im folgenden die vereinfachte Annahme getroffen, daß die Container einzeln bzw. nacheinander im KV–Terminal eintreffen. Ferner wird bei der Analyse das Vorhandensein zweier Portalkräne und zweier Lagerkräne, die folglich zwei Umschlagvorgänge gleichzeitig durchführen können, dadurch berücksichtigt, daß die Ankunftsrate der Container halbiert wird. Bei der Analyse des Modells wird folglich nur ein Portalkran und ein Lagerkran berücksichtigt. Bei einer Ankunftsrate von N Zügen pro Tag wird bei der Analyse somit eine Ankunftsrate von $30N/2$ Containern pro Tag angenommen. Schließlich wird in dem Modell angenommen, daß die zeitlichen Schwankungen der Ankunftsabstände der Container durch den Variationskoeffizienten $\sqrt{0.1}$ erfaßt werden. Somit läßt sich die Analyse des obigen Modells anhand eines einfachen Upper–Bound Modells mit zwei parallelen Bedienern durchführen. Die Bediener repräsentieren dementsprechend den Portalkran und den Lagerkran. Die Tatsache, daß die Kräne auf die Bereitstellung eines Containers auf der Vorsortierfläche durch den jeweiligen Gegenpart warten, wird dadurch berücksichtigt, daß die Schranken $U_{1,2}$ und $U_{2,1}$ des Upper–Bound Modells auf den Wert 1 gesetzt werden. Gemäß der Definition des Upper–Bound Modells ist durch diese Schrankenwahl gesichert, daß der Bediener, der seinen Teilauftrag zuerst bearbeitet hat, blockiert und auf den jeweils anderen Bediener wartet. Der vorliegende Fall ist somit ein Beispiel dafür, daß das Upper–Bound Modell das real erwünschte Systemverhalten exakt abbildet.

Anhand der Analyse des Upper–Bound Modells wurde abhängig von der Ankunftsrate der Container die mittlere Anzahl Container im System ermittelt. Dabei wurde davon ausgegangen, daß das KV–Terminal täglich für 10 Stunden betrieben wird. Da gemäß obiger Ausführungen Wartezeiten vermieden werden sollten, wurde die maximale Ankunftsrate ermittelt, bei der die mittlere Anzahl der Container im System höchstens den Wert 1 aufweist. Die für die Container gewonnenen Resultate wurden auf die Züge übertragen und sind anhand der Kurve in der Abbildung 8.5 veranschaulicht.

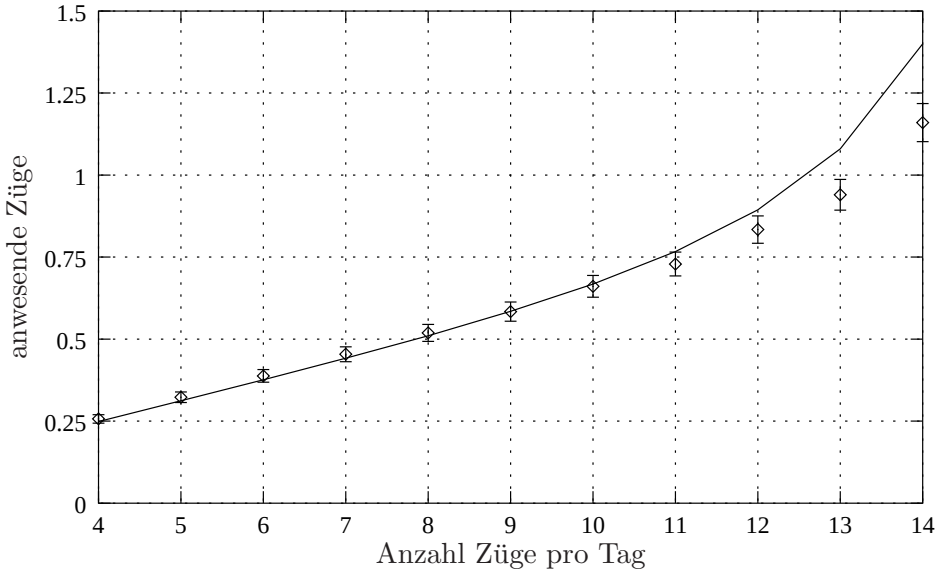

Abbildung 8.5: Anzahl anwesender Züge in Abhängigkeit der Ankunftsrate

Das Ergebnis der Analyse besagt folglich, daß anhand des modellierten KV–Terminals unter der oben geschilderten Nebenbedingung täglich 12 Züge umgeschlagen werden können. Bei angenommenen 240 Betriebstagen entspricht dies einer Jahresleistung von 86.400 Containern.

Zusätzlich zu den Analyseresultaten werden im folgenden die Aspekte Analyseaufwand und Approximationsgüte betrachtet. Der Analyseaufwand hinsichtlich Rechenzeit und Speicherplatzbedarf ergibt sich maßgeblich aus der Größe der involvierten QBDs bzw. der entsprechenden Matrizen. Diese Größe wiederum resultiert aus dem Produkt der Kardinalität der Makrozustandsräume $\tilde{Z}_k$ des Upper–Bound Modells (vgl. Gleichung (5.9)) und der Phasenanzahlen der Zwischenankunftszeit– und Bedienzeitverteilungen. Mit den Schranken $U_{1,2} = U_{2,1} = 1$ besitzen die $\tilde{Z}_k$ 3 Makrozustände. Aufgrund der im Modell gewählten Variationskoeffizienten besitzt die Zwischenankunftszeitverteilung 10 Phasen und die Bedienzeitverteilungen je 5 Phasen. Somit besitzen die bei der Analyse betrachteten QBDs 750 Zustände. Auf einem AMD Opteron Prozessor mit 2 GHz, 1 MB Cache und 6 GB Hauptspeicher ließen sich diese QBDs je Analyselauf in weniger als einer Minute analysieren.

Zur Überprüfung der Approximationsgüte der Analyseresultate wurden die in der Abbildung 8.5 dargestellten mittleren Zuganzahlen mit Simulationsergebnissen für das in der Abbildung 8.4 dargestellte ProC/B–Modell verglichen. Die Simulationsergebnisse wurden jeweils nach Erreichen eines 90% Konfidenzintervalls der Breite $\leq 5\%$ abgebrochen. Die Simulationsergebnisse sowie die zugehörigen 90% Konfidenzintervalle sind ebenfalls in der Abbildung 8.5 anhand der Fehlerbalken dargestellt. Es zeigt sich, daß die Approximationsgüte in den Fällen $N = 1, \ldots, 12$ sehr hoch ist. Lediglich in den Fällen $N > 12$, in denen die mittlere Anzahl der Container bzw. Züge des Upper–Bound Modells den Wert 1 übersteigt, weisen die Resultate größere Abweichungen von den Simulationsergebnissen auf. Um den Ursachen für diese Abweichungen nachzugehen, seien nochmals die zur Analyse des Upper–Bound Modells getroffenen Vereinfachungen betrachtet. Die erste Vereinfachung liegt in der Berücksichtigung zweier Portalkräne und zweier Lagerkräne durch Halbierung der Ankunftsrate. Da auf diese Weise ein paralleler Ablauf, nämlich die parallele Tätigkeit zweier Portalkran–Lagerkran Paare aus dem Modell herausgenommen wird, führt diese Vereinfachung eher zu einer Unterschätzung der Resultate für das reale System. Die Ursache für die Approximationsfehler im Fall $N > 12$ liegt folglich in der Approximation der Batch–Ankünfte durch Einzelankünfte der Container. Im Fall der im Simulationsmodell betrachteten Batch–Ankünfte beginnen die Portalkräne und Lagerkräne nach Beendigung eines Umschlagvorgangs unmittelbar mit dem nächsten Umschlagvorgang, solange nicht sämtliche Container eines Zuges verbracht worden sind. Dies ist jedoch in der Analyse des Upper–Bound Modells, in dem die Container einzeln ankommen, nicht der Fall. Dadurch entstehen immer wieder künstlich in das Modell induzierte Wartezeiten der Portalkräne und der Lagerkräne. Aufgrund dieser künstlichen Wartezeiten wird die von dem System täglich zu bewältigende Last eher unterschätzt.

Zusammenfassend erweist sich die vorgestellte Methode zur Grobplanung von Erweiterungs– bzw. Umbaumaßnahmen für KV–Terminals als geeignet. Insbesondere läßt sich die von dem System zu bewältigende Containerlast abschätzen. Ferner lassen sich mit Little's Gesetz aus den in der Abbildung 8.5 dargestellten Resultate die mittleren Durchlaufzeiten bzw. Umschlagzeiten der Züge in dem KV–Terminal ermitteln. Die Umschlagzeiten sind in der Abbildung 8.6 wiederum in Abhängigkeit der Ankunftsrate der Züge dargestellt. Ebenso sind die 5% breiten 90%-Konfidenzintervall der Simulationsergebnisse durch die Fehlerbalken.

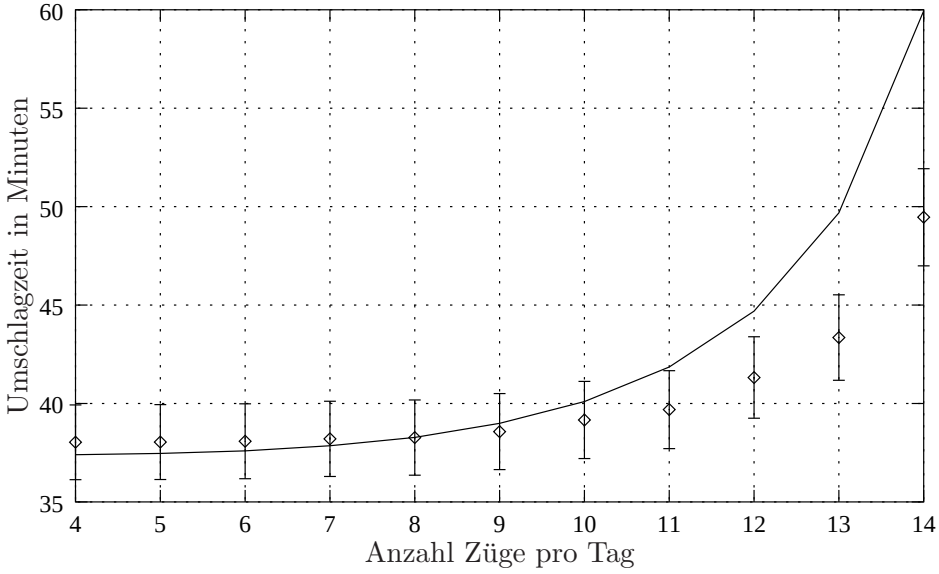

Abbildung 8.6: Verweilzeit der Züge in Abhängigkeit der Ankunftsrate

Im nächsten Abschnitt wird gezeigt, wie die in dieser Arbeit entwickelte Analysemethode für erweiterte Fork/Join–Warteschlangennetze genutzt werden kann, um Zulieferprozesse in der Automobilindustrie zu planen.

8.3 Lieferketten in der Automobilindustrie

Die Automobilindustrie ist in hohem Maße Spiegelbild wirtschaftlicher Entwicklungen. An ihr lassen sich die Verschärfung des Wettbewerbs, Auswirkungen der Globalisierung, verändertes Kundenverhalten und Konjunktur–bedingte Nachfrageschwankungen deutlich ablesen. Die optimale Planung, Steuerung und Kontrolle aller betrieblichen und logistischen Prozesse ist somit besonders bedeutsam und entscheidet über Erfolg und Mißerfolg. Diese Aussage läßt sich leicht durch die Anzahl weltweit agierender Automobilhersteller belegen, die von 50 Unternehmen im Jahr 1964 auf nur noch 12 Unternehmen im Jahr 2000 abnahm.

Somit gilt die in der Einführung zu diesem Kapitel erwähnte Notwendigkeit moderner Industrieunternehmen, sich auf Kernkompetenzen zu konzentrieren und Kooperationen mit Unternehmenspartnern einzugehen, für Automobilhersteller in besonderer Weise. Ein Beleg dafür ist die Tatsache, daß die Wertschöpfungstiefe eines Automobilherstellers i.a. unter 25% liegt. Verstärkt wird diese Entwicklung durch die extrem hohe Produktkomplexität, durch technologische Innovation und stark differierende Kundenanforderungen (vgl. [97]). Für die Automobilindustrie resultieren daraus spezifische Herausforderungen wie eine zunehmende Variantenvielfalt, sinkende Produktlebenszyklen, steigende Produktkomplexität und kürzere Lieferzeiten. Damit verbunden sind längere Entwicklungszeiten und höhere Entwicklungskosten, längere Amortisationszeiten und kürzere Zeitfenster zur Gewinnerzielung. Um diesen hohen Anforderungen gerecht zu werden und die steigende Prozeßkomplexität bewältigen zu können, werden in zunehmendem Maße Verantwortungsbereiche sowohl in der Produktentwicklung als auch in der Produktion auf die Zulieferer übertragen. Mit diesem Trend ist konsequenterweise ein Veränderungsprozeß in der Zulieferindustrie verbunden. Aufgrund des gestiegenen Verantwortungsbereichs und der stärkeren Integration in den Herstellungsprozeß verändert sich die Rolle eines Zulieferers vom reinen Teile–Lieferanten hin zum Modul–Lieferanten. Modul–Lieferanten liefern komplette Teilsysteme wie Armaturentafeln, Inneneinrichtungen, Bremssysteme, Fahrwerke etc. und montieren diese teilweise zudem beim Hersteller. Weiterhin gelten die Aussagen hinsichtlich der Wettbewerbssituation und der Konzentration auf Kernkompetenzen unverändert auch für die Modul–Lieferanten. Daraus ergibt sich die in der Abbildung 8.7 dargestellte mehrstufige Lieferkette.

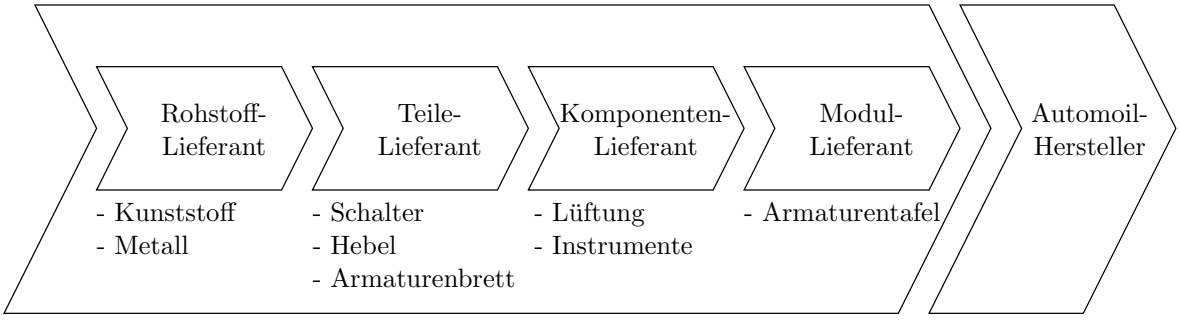

Abbildung 8.7: Zulieferkette in der Automobilindustrie

Die optimale Planung, Steuerung und Kontrolle der Logistik und speziell der Kooperationstätigkeiten des Automobilherstellers mit seinen Modul–Lieferanten stellt folglich ein wichtiges Kriterium für die Wettbewerbsfähigkeit und den Erfolg aller an einer Kooperation betei-

ligten Unternehmen dar. In diesem Kapitel wird exemplarisch die Planung des Lieferprozesses eines Armaturentafel–Lieferanten erläutert. Dazu wird im folgenden zunächst die Problemstellung konkretisiert und anhand eines ProC/B–Modells abgebildet. Anschließend wird im Abschnitt 8.3.2 das Modell anhand des Analyseverfahrens für Fork/Join–Warteschlangennetze analysiert.

8.3.1 Modellbeschreibung

Betrachtet wird die Planung des Zulieferprozesses eines Armaturentafel–Lieferanten. Die Aufgabe des Lieferanten besteht darin, die Armaturentafeln herzustellen und Just–in–Sequence an den Montageort des Automobilherstellers zu liefern. Die Planung des Zulieferprozesses unterliegt gewissen Rahmenbedingungen, die einerseits durch Absprachen bzw. Verträge mit dem Automobilhersteller geregelt sind und die sich andererseits aus dem Produktionsprozeß des Zulieferers ergeben. Ferner verbleiben gewisse Freiheitsgrade, deren optimale Ausgestaltung z.B. hinsichtlich der entstehenden Kosten im Interesse des Zulieferers liegt. Im folgenden werden zunächst diese Rahmenbedingungen sowie die Freiheitsgrade, die in der Hand des Zulieferers liegen, erläutert. Anschließend wird ein ProC/B–Modell vorgestellt und analysiert, um dem Zulieferer eine Entscheidungshilfe bei der Belegung der Freiheitsgrade zu geben.

Der Automobilhersteller fertigt täglich in einer Mischform aus Fließfertigung und Gruppenfertigung etwa 400 PKW und benötigt folglich ebenso viele Armaturentafeln. Die Fertigung ist in zwei Schichten zu je 8 Stunden organisiert, so daß etwa alle 2,4 Minuten eine Armaturentafel benötigt wird. Die Aufgabe des Armaturentafel–Lieferanten besteht darin, die Armaturentafeln Just–in–Sequence an den Montageort des Herstellers zu liefern. Um Lieferengpässe des Armaturentafel–Lieferanten und damit einen Verzug in der Produktion des Herstellers zu vermeiden, ist es in der Praxis üblich, daß die Zulieferer in räumlicher Nähe zum Hersteller ein Lager unterhalten. Gewöhnlich verlangt der Hersteller, daß in dem Lager stets ein Mindestbestand von drei Produktionstagen vorhanden ist. Um diese Bedingung einzuhalten, muß der Zulieferer die Armaturentafeln folglich spätestens drei Tage vor dem Verbau an das Lager liefern. Andererseits besitzt der Zulieferer aufgrund der Produktionsplanung des Herstellers erst fünf Tage vor dem Verbau der Armaturentafeln exakte Informationen über Variantenspezifika und damit die konkrete Ausprägung der zu fertigenden Armaturentafeln. Folglich kann er für die Produktion und Anlieferung (Produktionszyklus) der Armaturentafeln maximal zwei Tage einplanen. Die Zeitpunkte, zu denen die Varianten–spezifischen Armaturentafeln benötigt werden, kann der Zulieferer in diesem Zeitraum aus einem Informationssystem des Automobilherstellers entnehmen.

Der Freiheitsgrad des Armaturentafel–Lieferanten, der im folgenden unter diesen Rahmenbedingungen betrachtet wird, liegt in der Wahl der Losgröße L bzw. in der Wahl der Anzahl N täglicher Belieferungen. Die Losgröße L und die Anzahl täglicher Belieferungen stehen offensichtlich in folgendem Zusammenhang:

$$N = \frac{400}{L} \tag{8.1}$$

Die Bestimmung der optimalen Losgröße ist eine klassische Aufgabe der Logistik und hat wesentlichen Einfluß auf die Kosten, die dem Zulieferer aus seinen Aktivitäten entstehen. Die auftretenden Kosten werden in losfixe und losvariable Kosten unterschieden. Losfixe Kosten sind unabhängig von der Losgröße und entstehen z.B. aus Auftragsbearbeitungen und Maschinen–Rüstzeiten. Um die losfixen Kosten gering zu halten, sollte offensichtlich eine geringe Anzahl N täglicher Belieferungen gewählt werden. Andererseits hängen losvariable Kosten wie z.B. Lagerzeiten von der Losgröße ab. Um hohe Lagerkosten zu vermeiden, sollte die Losgröße möglichst klein gewählt werden und somit die Anzahl täglicher Belieferungen eher groß. Um diesem Zielkonflikt einer optimalen Losgröße zu begegnen, existieren in der Logistik verschiedenen Verfahren, sog. Losgrößenmodelle (vgl. [51]). Auf diese Problematik wird im nachfolgenden Abschnitt 8.3.2 genauer eingegangen.

Neben dem Kostenfaktor hat die Losgröße wesentlichen Einfluß auf die Dauer des Produktionszyklus eines Fertigungsloses. Die Kenntnis des Produktionszyklus ist für den Zulieferer essentiell, da sich daraus und aus dem vorgegebenen Lieferzeitpunkt der Produktionsbeginn eines Fertigungsloses zurückrechnen läßt. Im folgenden wird zunächst das Modell vorgestellt, das dem Armaturentafel–Lieferanten erlauben wird, in Abhängigkeit der Losgröße die Dauer eines Produktionszyklus bzw. den Zeitpunkt des Produktionsbeginns eines Fertigungsloses zu bestimmen. Das Modell ist hierarchisch strukturiert und in den Abbildungen 8.8 bis 8.11 dargestellt.

Die Abbildung 8.8 veranschaulicht die oberste Hierarchiestufe des Modells aus Sicht des Automobilherstellers. Ebenso wie der Hersteller arbeitet der Zulieferer in zwei Schichten zu je 8 Stunden, so daß alle $\frac{16 \cdot 60}{400/L}$ Minuten die Produktion eines Fertigungsloses beginnt. Diese Tatsache ist entsprechend in der Quelle des ProC/B–Modells berücksichtigt. Ein Prozeß der dargestellten Prozeßkette korrespondiert jeweils zu einem Produktionszyklus, d.h. zu der Produktion und Anlieferung eines Fertigungsloses des Armaturentafel–Lieferanten. Die in dem dargestellten Prozeßkettenelement angedeutete Fertigung und Anlieferung der Armaturentafeln (AT) wird durch die Funktionseinheit *AT_Zulieferer* in Abbildung 8.9 verfeinert.

Die Abbildung 8.9 stellt die wesentlichen Schritte zur Fertigung der Armaturentafeln anhand der Prozeßkettenelemente dar. In der Fertigung der Armaturentafeln beschränkt sich der Lieferant im wesentlichen auf das Zusammensetzen einzelner Komponenten, die er wiederum von Komponenten–Lieferanten bezieht. Die entsprechenden Lieferaufträge werden von der Administration je Fertigungslos erstellt. In dem Modell werden zwei Komponenten–Lieferanten betrachtet, die zum einen Instrumente wie z.B. Tachometer und Drehzahlmesser sowie Hebel fertigen und zum anderen die Armaturenbretter (AB) herstellen. Aus Kostengründen wird an dieser Schnittstelle auf ein Lager zur Zwischenlagerung der Komponenten verzichtet. Die Fertigung der Armaturentafeln beginnt unmittelbar nach der Belieferung durch die beiden Komponenten–Lieferanten. Die weiteren Arbeitsschritte bestehen in der Prüfung, ggfls. Nachbesserung und Verpackung der Armaturentafeln. Schließlich wird das Fertigungslos an den Automobilhersteller ausgeliefert.

Die Tätigkeiten der Komponenten–Lieferanten sind in den Abbildungen 8.10 und 8.11 dargestellt. Auch im Falle der Komponenten–Lieferanten erfolgt zunächst eine Bearbeitung der Aufträge. Dabei werden z.B. benötigte Teile oder Rohmaterialien reserviert und Nachbestellungen veranlaßt. Anschließt erfolgt die parallele Produktion und Prüfung verschiedener

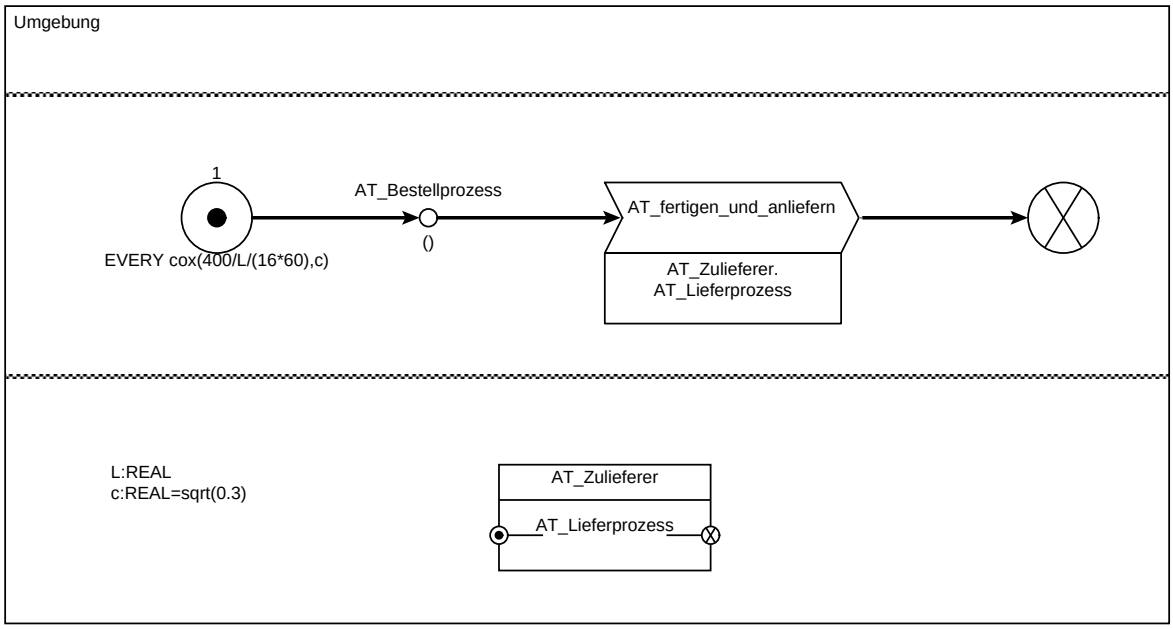

Abbildung 8.8: Umgebung des Armaturentafel–Lieferanten

Komponenten, bevor diese ggfls. nachgebessert und anschließend verpackt werden. Zuletzt werden die Komponenten an den Armaturentafel–Lieferanten geliefert.

Mit diesen Ausführungen ist das Modell vollständig beschrieben und wird im folgenden Abschnitt analysiert.

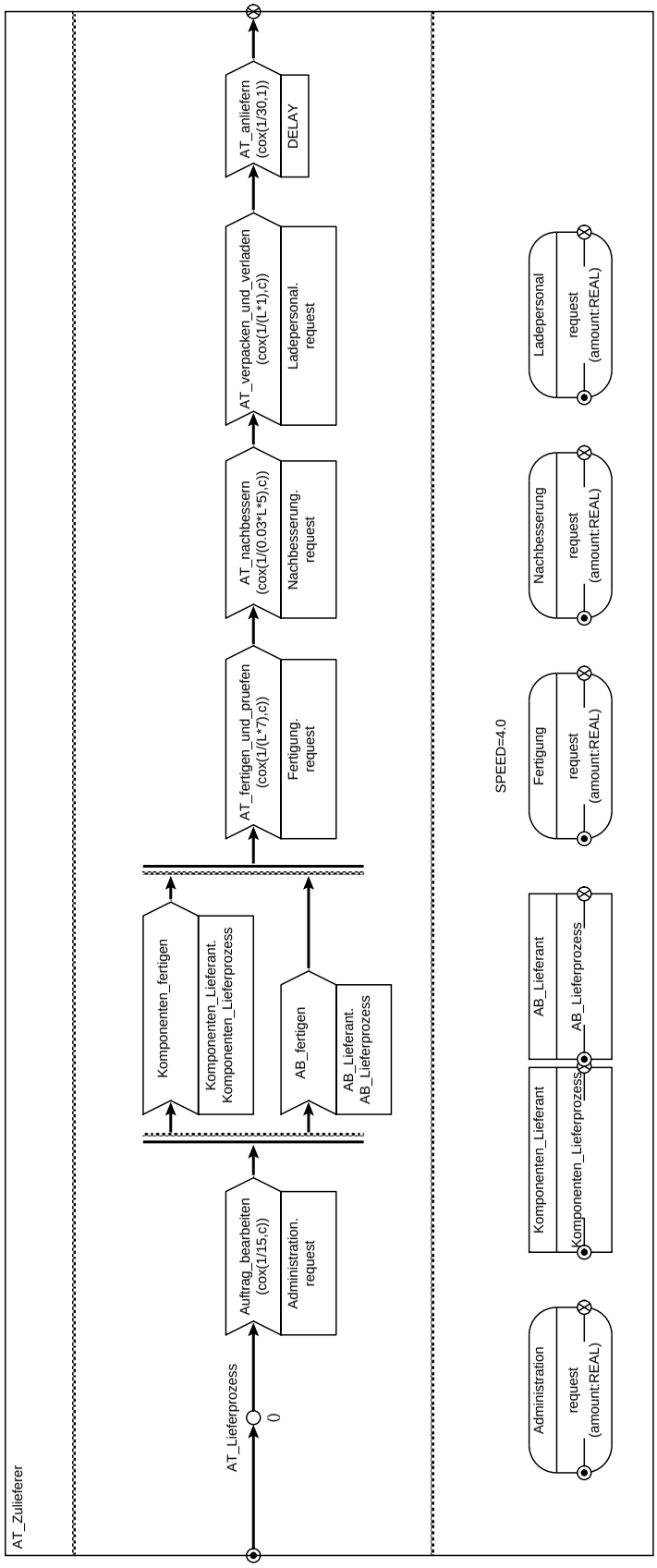

Abbildung 8.9: Armaturentafel–Lieferant

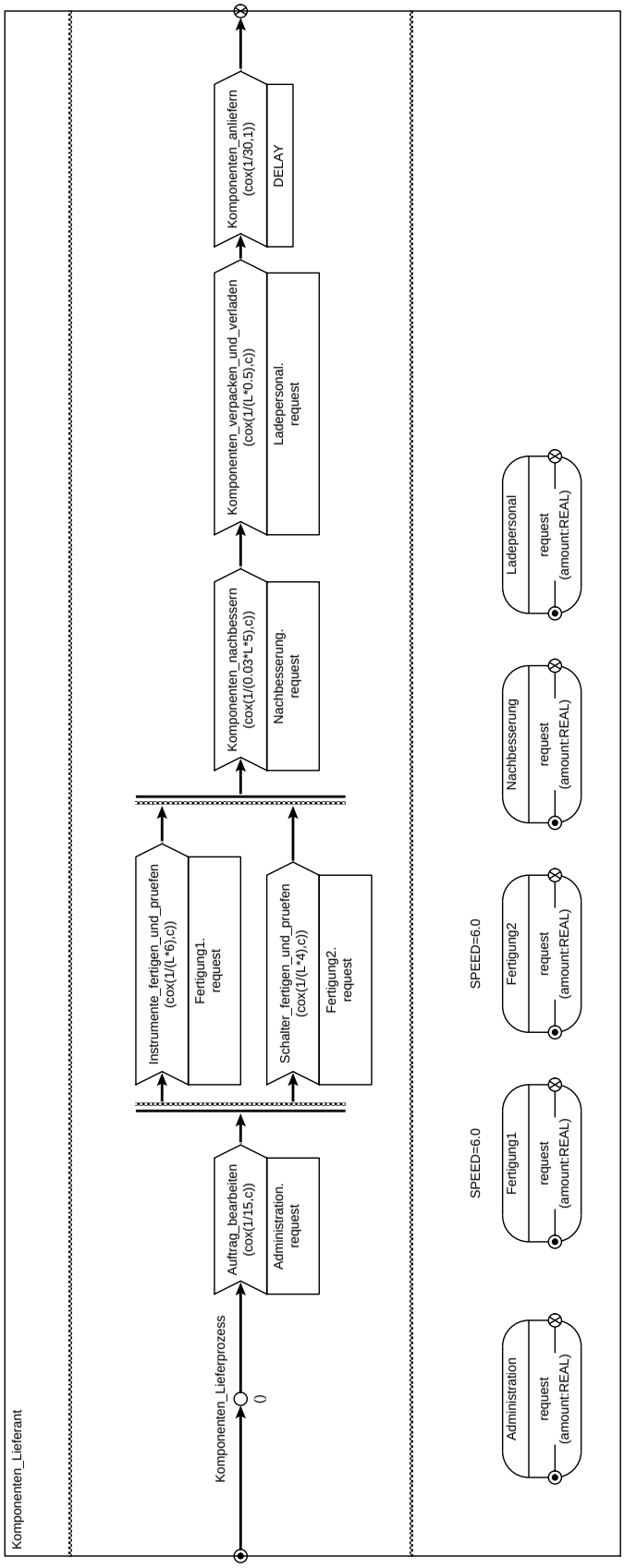

Abbildung 8.10: Komponenten–Lieferant

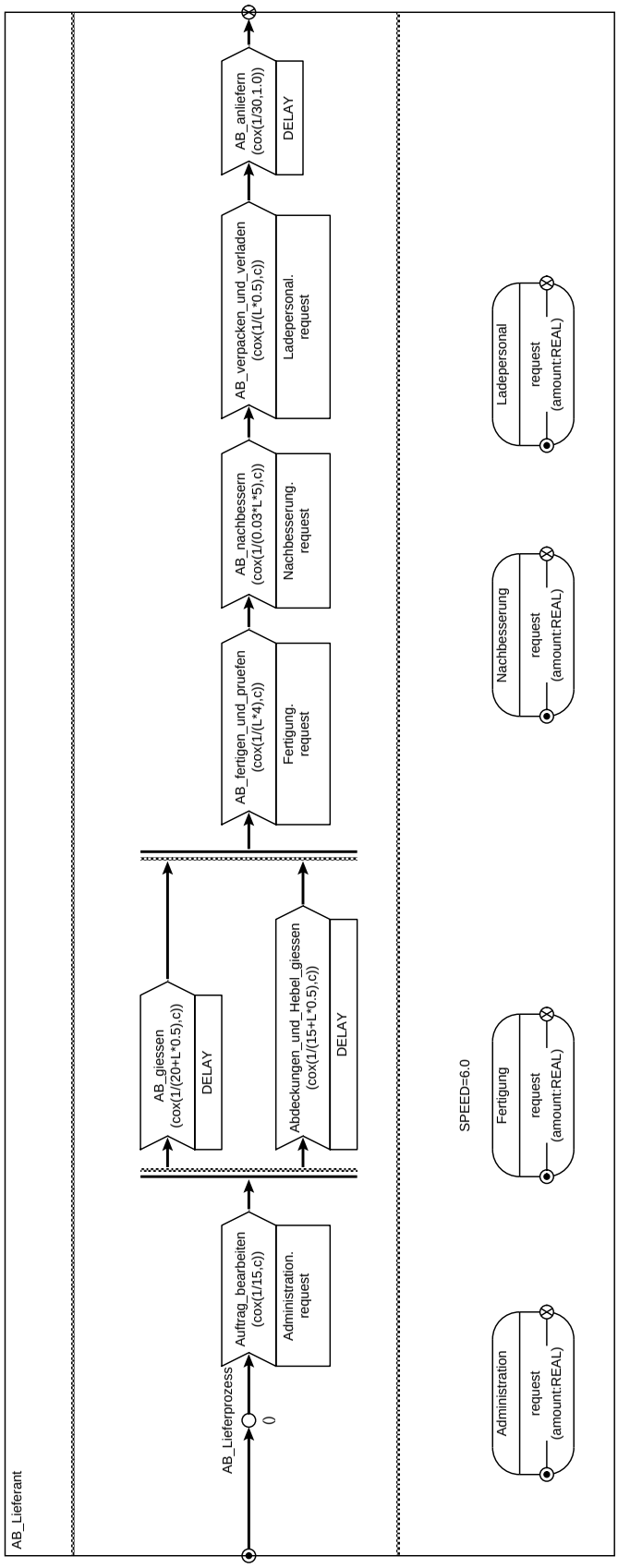

Abbildung 8.11: Armaturenbrett–Lieferant

8.3.2 Analyse

Das Interesse an der Analyse des beschriebenen Modells liegt in der Bestimmung der Dauer eines Produktionszyklus in Abhängigkeit der Losgröße L bzw. der Anzahl täglicher Belieferungen N. Konkret wurde das Modell für die Fälle $N = 1, \ldots, 10$ betrachtet (die Parameter L des Modells ergeben sich aus Gleichung (8.1)).

Bevor die erzielten Analyseresultate dargestellt werden, wird im folgenden zunächst auf die Anwendung des Analyseverfahrens für erweiterte Fork/Join–Warteschlangennetze auf das vorliegende Modell eingegangen. Gemäß der hierarchischen Modellstruktur erfolgt die Analyse ebenfalls mehrstufig. Zunächst werden die Teilmodelle auf unterster Hierarchiestufe, also die ProC/B–Modelle der Abbildungen 8.10 und 8.11 betrachtet. Zur Analyse dieser Modelle ist zunächst das Verhalten der Quelle, d.h. die Verteilungen der Ankunftsprozesse zu bestimmen. Diese ergeben sich aus der Verteilung der Zwischenabgangszeiten der Administration des Armaturentafel–Lieferanten aus Abbildung 8.9. Damit läßt sich das Modell des Komponenten–Lieferanten unmittelbar analysieren. Im Fall des Armaturenbrett–Lieferanten sind zunächst Aggregate für die Prozeßkettenelemente *AB_giessen* und *Abdeckungen_und_Hebel_giessen* des ProC/B–Modells aus Abbildung 8.11 zu bestimmen. Durch Einsetzen dieser Aggregate in das Fork/Join–Teilmodell läßt sich damit auch das ProC/B–Modell des Armaturenbrett-Lieferanten analysieren. Nach diesem Schritt liegen Analyseresultate für die ProC/B–Modelle des Komponenten–Lieferanten und des Armaturenbrett–Lieferanten vor. Insbesondere sind die entsprechenden ersten beiden Momente der Durchlaufzeitverteilungen bekannt. Auf dieser Grundlage werden im nächsten Schritt für beide Teilmodelle $\mathcal{H}/\mathcal{H}/1/\infty$ –Aggregate bestimmt. Diese Aggregate werden schließlich in das Fork/Join–Teilnetz des ProC/B–Modells aus Abbildung 8.9 eingesetzt. Somit läßt sich ebenfalls mithilfe des Analyseverfahrens für erweiterte Fork/Join–Warteschlangennetze die Durchlaufzeit des Modells für den Armaturentafel–Lieferanten ermitteln. Anhand dieser Durchlaufzeit kann der Zulieferer somit den Produktionsbeginn eines Fertigungsloses bestimmen, um rechtzeitig drei Tage vor dem Verbau der Armaturentafeln das Lager beliefern zu können. Die Ergebnisse für $N = 1, \ldots, 10$ tägliche Lieferungen sind in der Abbildung 8.12 dargestellt.

Aus den Rahmenbedingungen, die dem Lieferprozeß zugrundeliegen, ging hervor, daß der Armaturentafel–Zulieferer maximal zwei Arbeitstage für einen Produktionszyklus inklusive Transportzeiten zur Verfügung hat. Diese Bedingung resultierte einerseits aus der Anforderung eines minimalen Lagerbestandes von drei Produktionstagen des Automobilherstellers und andererseits aus der Produktionsplanung des Herstellers, aus der exakte Informationen über Variantenspezifika frühestens fünf Arbeitstage vor dem Verbau der Armaturentafeln zur Verfügung stehen. Unter den gegebenen Umständen entsprechen zwei Arbeitstage zu je zwei Schichten $2 \cdot 16 \cdot 60$ Minuten = 1920 Minuten. Aus der Abbildung ist leicht ersichtlich, daß bei einer einzigen Belieferung pro Tag (und einer Losgröße $L = 400$) die Bedingung eines Produktionszyklus von maximal zwei Tagen nicht einzuhalten ist. In allen übrigen Fällen liegt ein Produktionszyklus inklusive Transportzeit jeweils unter zwei Tagen.

Zur Beurteilung der Approximationsgüte wurden die Analyseresultate mit je einer Simulation des Modells für $N = 1, \ldots, 10$ verglichen. Die Simulationsläufe wurden jeweils nach Erreichen eines 90%–Konfidenzintervalls der Breite $\leq 5\%$ beendet. Die Simulationsergebnisse sind ebenfalls in der Abbildung 8.12 anhand der Fehlerbalken dargestellt. Offensichtlich weisen

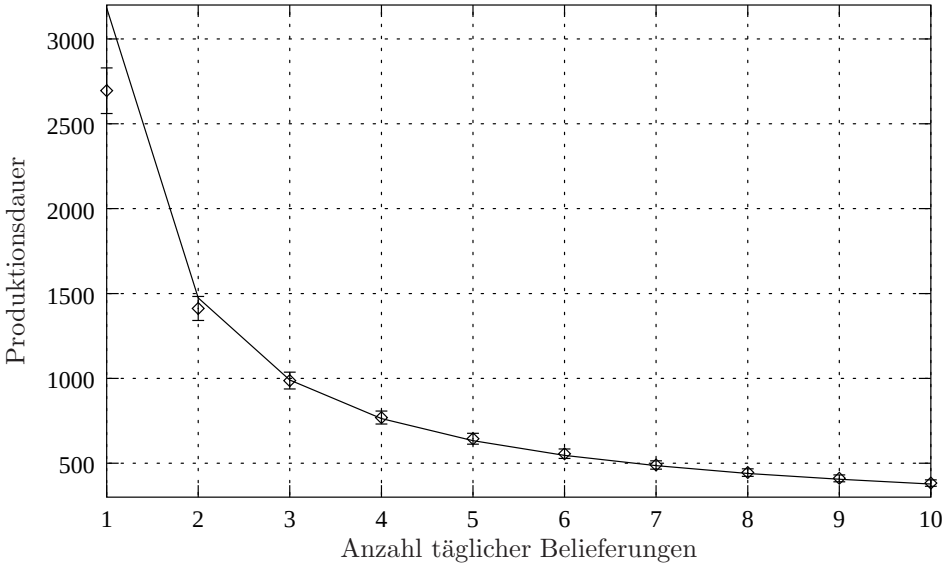

Abbildung 8.12: Produktionsdauer des Armaturentafel–Lieferanten

die Resultate des Verfahrens für erweiterte Fork/Join–Warteschlangennetze für das vorliegende Modell eine sehr hohe Approximationsgüte auf. Lediglich bei einer einzigen täglichen Belieferung ist dies nicht der Fall. Die Begründung hierfür liegt in der im Fall $N = 1$ recht hohen Auslastung der parallelen Bediener der im Modell enthaltenen Fork/Join–Teilmodelle. In den Experimenten in Abschnitt 5.2 wurde bereits vermutet, daß die Approximationsgüte des Upper–Bound Modells mit steigender Auslastung der parallelen Bediener abnimmt. Ferner hängt die Approximationsgüte maßgeblich von der Schranke ϵ des Upper–Bound Modells ab (vgl. Gleichung (5.57) in Abschnitt 5.1.4). Diese Schranke gibt die Wahrscheinlichkeit dafür an, daß in dem Upper–Bound Modell einer der parallelen Bediener durch den jeweils anderen blockiert wird. In den Untersuchungen an dem vorliegenden Modell wurde aus Aufwandsgründen $\epsilon = 0.025$ gewählt. Im Vergleich dazu lag diese Schranke in den Experimenten des Abschnitts 5.2 bei 0.005, so daß im vorliegenden Modell mit höheren Approximationsfehlern als in Abschnitt 5.2 zu rechnen ist. Eine deutliche Verringerung von ϵ würde jedoch aufgrund der recht kleinen Variationskoeffizienten der Ankunfts– und Bedienprozesse der Fork/Join–Teilmodelle zu unakzeptablen Rechenzeiten des Analyseverfahrens führen.

Der Berechnungsaufwand und der Speicherplatzbedarf des Verfahrens wird im wesentlichen von den Fork/Join–Strukturen in den Teilmodellen (vgl. Abb. 8.9- 8.11) bestimmt. Die Zustandsraumgrößen zur Analyse der Fork/Join-Teilnetze in den Teilmodellen 8.10 und 8.11 lagen jeweils etwa bei 320 Zuständen. Die Zustandsraumgröße zur Analyse des Fork/Join-Teilnetzes in Abb. 8.9 lag jeweils bei etwa 600-800 Zuständen. Derartige Zustandsräume lassen sich anhand Matrix–geometrischer Verfahrens mit recht geringem Rechenaufwand behandeln. Auf einem AMD Opteron Prozessor mit 2 GHz, 1 MB Cache und 6 GB Hauptspeicher lag die Antwortzeit des Verfahrens zur Erzielung der oben dargestellten Analyseresultate bei etwa 2 Minuten.

Aus den gewonnenen Analyseresultaten kann der Zulieferer schließlich eine optimale Losgröße sowie den Produktionsbeginn eines Fertigungsloses bestimmen. Zur Berechnung der optimalen

Losgröße eignet sich im vorliegenden Fall das klassische Losgrößenmodell nach Harris und Andler (vgl. [51]). Dazu seien:

- K die losfixen Produktionskosten eines Fertigungsloses,
- c die variablen Produktionskosten je Mengeneinheit,
- h die Lagerkosten je Mengen– und Zeiteinheit,
- $s(L)$ die von der Losgröße abhängenden Transportkosten.

Ferner ergibt sich die mittlere Lagerzeit $T(L)$ eines Fertigungsloses der Größe L zu:

$$T(L) = 3 \cdot 960 + \frac{1}{2} L \frac{960}{400}.$$

Damit resultieren je Fertigungslos in Abhängigkeit der Losgröße L die folgenden Kosten $C(L)$:

$$C(L) = K + cL + hLT(L) + s(L) \tag{8.2}$$

Die Multiplikation der Gleichung 8.2 mit $N = 400/L$ liefert die täglichen Gesamtkosten $G(L)$ des Armaturentafel–Lieferanten. Es gilt:

$$G(L) = 400\frac{K}{L} + 400c + 400hT(L) + \frac{400}{L}s(L) \tag{8.3}$$

Aus $G'(L^*) = 0$ läßt sich schließlich unter der Kenntnis von $s(L)$ die optimale Losgröße L^* ermitteln.

Somit leistet das Analyseverfahren für erweiterte Fork/Join–Warteschlangennetze im Fall des Armaturentafel–Lieferanten einen wichtigen Beitrag hinsichtlich der Optimierung der Losgröße. Zusammenfassend belegen die beiden Beispiele aus dem Bereich Logistik den Nutzen des vorgestellten Analyseverfahrens für die Grobplanung und Optimierung logistischer Netzwerke.

Kapitel 9

Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Verfahren zur Modellierung und Analyse von erweiterten Fork/Join–Warteschlangennetzen vorgestellt. Erweiterte Fork/Join–Warteschlangennetze eignen sich zur Modellierung der Synchronisation paralleler Abläufe und sind somit für viele Anwendungsbereiche von wichtiger Bedeutung. Sie können z.B. zur Unterstützung der Planungs– und Optimierungsprozesse von Computer– und Kommunikationssystemen, Produktionsanlagen und Logistiknetzen genutzt werden.

Grundlage des vorgestellten Verfahrens bildet das Dekompositionsverfahren nach Kühn/Whitt, das im ersten Teil der Arbeit erläutert wurde. Das Dekompositionsverfahren erweist sich gegenüber alternativen Analyseverfahren für offene Warteschlangennetze als besonders flexibel, da es aus rein technischer Sicht die Analyse beliebiger Netze erlaubt, sofern geeignete Methoden zur Behandlung der isoliert betrachteten Stationen verfügbar sind. Die Verwendung von Phasenverteilungen zur Approximation der Zwischenankunfts– und Bedienzeiten ermöglicht die isolierte Analyse einer reichhaltigen Klasse von Stationstypen, deren Dynamik durch QBDs darstellbar ist.

Diese Eigenschaft wurde im zweiten Teil der Arbeit ausgenutzt, um sog. einfache Fork/Join–Stationen im Kontext des Dekompositionsverfahrens beherrschbar zu machen. Dazu war die Approximation durch ein von Balsamo et al. [11] vorgestelltes Upper–Bound Modell notwendig. Die Analyseresultate des Upper–Bound Modells weisen im Vergleich zu einer Simulation des ursprünglichen, isoliert betrachteten Fork/Join–Modells eine sehr hohe Approximationsgüte auf. Im Kontext eines Warteschlangenmodells lieferte das um die Analyse von Fork/Join–Stationen angereicherte Dekompositionsverfahren in verschiedenen Experimentreihen teilweise recht zufriedenstellende Resultate. Speziell in zyklischen Netzen traten jedoch recht große Approximationsfehler auf. Im zentralen Kapitel 6 wurden spezielle $\mathcal{H}/\mathcal{H}/1/\infty$ –Aggregate vorgestellt. Diese Aggregate werden derart bestimmt, daß sie unter einem vorgegebenen $\mathcal{H}$ verteilten Ankunftsprozeß denselben Erwartungswert ED und dieselbe Varianz VD der Durchlaufzeitverteilung liefern, wie ein zu aggregierendes Warteschlangennetz. Es konnte gezeigt werden, daß dies immer dann gelingt, wenn ED^2 in einem durch VD bestimmten Bereich liegt. Diese Aggregate erlaubten die Behandlung deutlich komplexerer Fork/Join–Stationen. Hinsichtlich der praktischen Relevanz ist damit gegenüber den einfa-

chen Fork/Join–Stationen eine erheblich größere Modellklasse analysierbar. Der praktische Nutzen der erarbeiteten Analysetechnik für erweiterte Fork/Join–Warteschlangennetze wurde anhand einiger Beispiel im dritten Teil der Arbeit belegt.

Nach dieser skizzenhaften Zusammenfassung der zentralen Schwerpunkte dieser Arbeit stellt sich abschließend die Frage, inwieweit die zu Beginn formulierten Ziele erreicht werden konnten. Die Ziele lagen in der Verfügbarmachung effizienter nicht–simulativer Analyseverfahren für das Anwendungsgebiet Logistik. Im Rahmen des Sonderforschungsbereichs „Modellierung großer Netze der Logistik“ *(www.sfb559.uni-dortmund.de)* wurden typische Systemeigenschaften von Logistiknetzen herausgearbeitet, die i.a. eine direkte Verwendung der in der Informatik seit langem bekannten effizienten Analyseverfahren für diskrete ereignisgesteuerte dynamische Systeme erschweren. Eine dieser typischen Systemeigenschaften ist die Synchronisation paralleler Abläufe, die in zahlreichen Anwendungsfällen der Logistik von Bedeutung ist. Aus dieser Motivation heraus beschäftigte sich diese Arbeit mit der Analyse erweiterter Fork/Join–Warteschlangennetze unter Verwendung des Dekompositionsverfahrens nach Kühn/Whitt. Die Frage des letztendlichen Nutzens dieser Technik speziell im Anwendungsgebiet Logistik ist differenziert zu betrachten und wird im folgenden diskutiert.

Die am häufigsten verwendete Analysemethode in der Logistik ist die ereignisgesteuerte Simulation. Aus Gründen der Akzeptanz eines Modells liegt das Interesse häufig in einer in hohem Maße detailgetreuen Nachbildung des betrachteten Systems. Aufgrund des hohen Detaillierungsgrades und der Größe der Modelle schließt sich die Verwendung effizienter nicht–simulativer Analysetechniken i.a. aus. Andererseits ist die Simulation mit erheblichen Nachteilen verbunden. Nicht nur die teilweise langen Simulationsläufe und die häufig fehlerhafte Interpretation der Simulationsergebnisse, die aus unzureichender Erfahrung im Umgang mit den Simulationsergebnissen resultiert, sind ein Beleg dafür. Auch die Modellierungsphase ist oft sehr zeitintensiv, da keine Möglichkeiten zur Überprüfung der funktionalen Korrektheit eines Modells existieren und Modellierungsfehlern somit in mühevoller Arbeit nachgegangen werden muß. Um diesen Schwächen der Simulation entgegenzuwirken, erwuchs innerhalb des Sonderforschungsbereichs der Vorschlag, Logistiknetze in einem mehrstufigen Verfahren zu modellieren und zu analysieren (vgl. [4]). Es empfiehlt sich, zunächst recht grobe Modelle zu betrachten und diese schrittweise zu verfeinern. Besonders in frühen Modellierungsphasen, in denen Systeme auf einem recht hohen Abstraktionsniveau betrachtet werden, sind Warteschlangennetze häufig ein adäquates Beschreibungsinstrument. Ihre Analyse gewährt erste Einblicke in das betrachtete System und erlaubt die Ermittlung einiger im Logistikbereich interessierender Kennzahlen wie Auslastungen, Lieferzeiten erreichbare Durchsätze und beantwortet Fragen zur Dimensionierung von Ressourcen. In diesem Kontext liefert die vorliegende Arbeit einen wichtigen Beitrag, da sie das Potential der Analyseverfahren für Warteschlangennetze erweitert und hinsichtlich des im Logistikkontext essentiellen Aspektes der Synchronisation paralleler Abläufe anpaßt. Insbesondere die Verwendung der vorgestellten Aggregierungstechnik erlaubt die Analyse komplexer Fork/Join–Strukturen. Die Anwendungsfälle im dritten Teil belegen die Nützlichkeit und die Relevanz des Verfahrens zur Analyse von Logistik–Modellen. Darüber hinaus ist das Dekompositionsverfahren recht flexibel, da es sich um weitere für die Logistik relevante Stationstypen erweitern läßt. Denkbar ist auch die hybride Analyse in Kombination mit alternativen Verfahren (vgl. [5]). Unbeantwortet bleibt in dieser Arbeit jedoch die Frage, inwieweit sich diese Analysetechnik in sehr großen Modellen bewährt. In diesem Fall ist zu erwarten, daß die Approximationsannahmen

deutlich stärkere Auswirkungen auf die Güte der Analyseresultate haben, als dies in den recht kleinen Modellen dieser Arbeit der Fall ist. Da das Approximationsverfahren aus technischer Sicht die Ermittlung der zweiten Momente z.B. der Durchlaufzeit ermöglicht, wäre es weiterhin wünschenswert, für diese eine höhere Approximationsgüte anzustreben. Da das Dekompositionsverfahren zur Ermittlung netzweiter Resultate die Unabhängigkeit der isolierten Stationen annimmt, sind die Resultate für die zweiten Momente i.a. sehr vage. Zur Berechnung des für die Logistik recht wichtigen Leistungskriterium der Liefertreue ist jedoch die Kenntnis der zweiten Momente der Durchlaufzeit essentiell. Insgesamt jedoch mag die Logistik aufgrund der Verfügbarkeit effizienter nicht–simulativer Analyseverfahren im Rahmen prozeßorientierter Modellierungsinstrumentarien wie z.B. des ProC/B–Toolsets, von diesen Techniken profitieren.

Nicht zuletzt soll an dieser Stelle der Beitrag herausgestellt werden, den diese Arbeit aus Sicht der Informatik hinsichtlich der Erweiterung von Analyseverfahren für Warteschlangennetze leistet. Zum einen wurde die Nützlichkeit des Dekompositionsverfahren hinsichtlich seiner Flexibilität durch die Integration von Fork/Join–Stationen herausgestellt. Andererseits haben jedoch verschiedene Experimente prinzipielle Defizite des Verfahrens unterstrichen, die bereits in zahlreichen Arbeiten zu Erweiterungen geführt haben. Diese Defizite liegen einerseits in der teilweise unzureichenden Approximation des Verkehrsflusses und andererseits in der Vernachlässigung von Abhängigkeiten unter den Analyseresultaten für die isolierten Stationen bei der Berechnung von netzweiten Leistungsmaßen wie z.B. Durchlaufzeiten. An einigen Experimenten wurde deutlich, daß diese Unzulänglichkeiten insbesondere in zyklischen Netzen zu fehlerhaften Resultaten führen. In azyklischen Netzen konnten in den durchgeführten Experimenten dagegen häufig sehr gute Ergebnisse erzielt werde. Die letzte Erkenntnis läßt sich jedoch nicht auf erheblich größere Modelle verallgemeinern, da auch in diesem Fall damit zu rechnen ist, daß sich Abhängigkeiten unter den Analyseresultaten für die isolierten Stationen deutlicher auswirken werden. Zur Verbesserung der Approximation des Verkehrsflusses, d.h. zur Darstellung nicht notwendigerweise unabhängig identisch verteilter Zwischenankunftszeiten sind in einigen Arbeiten bereits Erfolge durch die Verwendung von MAPs [42, 43] dokumentiert. Aus Sicht der Informatik liegt der Beitrag dieser Arbeit insbesondere in der Präsentation der vorgestellten Aggregierungstechnik, die die Analyse komplexer Fork/Join–Strukturen erlaubt. Zahlreiche Anwendungsfälle aus dem Bereich Computer– und Kommunikationssysteme belegen die Notwendigkeit der Betrachtung von komplexen Fork/Join–Strukturen. Anhand des hoch aktuellen Themas Web–basierter Informationsdienste wurde das Analyseverfahren demonstriert. Die vorgestellte Aggregierungstechnik weist jedoch eine weitere angenehme Eigenschaft auf. So läßt sie sich in gewissem Sinne verallgemeinern. Die Grundlage zur Entwicklung der Aggregierungstechnik bilden die in den Sätzen 6.1 und 6.2 bewiesenen Monotonieaussagen für die Momente der Durchlaufzeit einer FCFS–Station mit speziellen phasenverteilten Ankunfts– und Bedienprozessen. Diese Monotonieaussagen werden sich leicht auf weitere Stationstypen erweitern lassen, so daß anhand dieser Technik weitere Aggregattypen bestimmbar sind. Weiterhin ist zu vermuten, daß sich dieses Verfahren ebenfalls zur Anpassung höherer Momente oder auch weiterer Leistungsmaße eignet. In diesem Sinne bietet die vorgestellte Aggregierungstechnik ein reichhaltiges Spektrum weiterer Einsatzmöglichkeiten, die es zu untersuchen gilt.

Literaturverzeichnis

[1] I. F. Akyildiz. Die erweiterte parametrische Analyse für geschlossene Warteschlangennetze. In H. Beilner (Hrsg), *Messung, Modellierung und Bewertung von Rechensystemen, 3*, Informatik-Fachberichte 110, 170–185. Springer, 1985.

[2] M. Arns und F. Bause. An Instructive Example for Pitfalls in Simulation of Logistic Networks. In *ESS'2001, Simulation in Industry, 13th European Simulation Symposium and Exhibition*, 420–423. Marseilles, France, 2001.

[3] M. Arns, M. Eickhoff, M. Fischer, C. Tepper und M. Völker. New Features in the ProC/B–Toolset. Tools of the 2003 Illinois Int. Multiconference on Measurement, Modelling, and Evaluation of Computer-Communication Systems. Technical Report 781, Universität Dortmund, Fachbereich Informatik, 2003.

[4] M. Arns, M. Fischer und P. Kemper. Anwendung nicht–simulativer Techniken zur Analyse eines dezentralen Güterverkehrszentrums. Interner SFB 559 Bericht 03017, 2003.

[5] M. Arns, M. Fischer, P. Kemper und C. Tepper. Supply Chain Modelling and its Analytical Evaluation. *Journal of the Operational Research Society*, 53(8), 885–894, 2002.

[6] M. Arns, M. Fischer, H. Tatlitürk, C. Tepper und M. Völker. Modeling and Analysis Framework of Logistic Process Chains. In *Proc. of Joint Tool Session at PNPM/MMB/PAPM Conferences*, 56–61. Aachen, Germany, 2001.

[7] M. Arns, M. Fischer, C. Tepper und M. Völker. Visualization of Analysis Results in the ProC/B-Toolset. In *Proc. 1st Int. Conf. on Quantitative Evaluation of Systems (QEST 2004)*, 318–319. Enschede, Netherlands, 2004.

[8] F. Bacelli. Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. Technical Report 426, INRIA Rocquencourt.

[9] F. Bacelli, A. Makowski und A. Shwartz. The fork-join queue and related systems with synchronization constraints. *Advances in Applied Probability*, 21, 629–660, 1989.

[10] F. Bacelli, W. Massey und D. Towsley. Acyclic Fork-Join Queueing Networks. *Journal of the ACM*, 36(3), 615–642, 1989.

[11] S. Balsamo, L. Donatiello und N. van Dijk. Bound performance models of heterogeneous parallel processing systems. *IEEE Transactions on Parallel and Distributed Systems*, 9(10), 1041–1056, 1998.

[12] Y. Bard. Some Extensions to Multiclass Queueing Network Analysis. In *Proc. 3rd Int. Sym. Modelling and Performance Evaluation of Computer Systems*, 1, 51–62. North-Holland, 1979.

[13] F. Baskett, K. Chandy, R. Muntz und F. Palacios. Open, Closed and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22(2), 248–260, 1975.

[14] F. Bause. Queueing Petri Nets - A Formalism for the Combined Qualitative and Quantitative Analysis of Systems. In *5th Int. Workshop on Petri Nets and Performance Models*, 14–23. IEEE Press, Toulouse, France, 1993.

[15] F. Bause, P. Buchholz und P. Kemper. A toolbox for functional and quantitative analysis of DEDS. In *Quantitative Evaluation of Computer and Communication Systems*, LNCS 1469, 356–359. Springer, 1998.

[16] F. Bause, P. Kemper und P. Kritzinger. Abstract Petri Net Notation. *Petri Net Newsletter*, 49, 9–27, 1995.

[17] B. Baynat und Y. Dallery. A unified view of product-form approximation techniques for general closed queueing networks. *Performance Evaluation*, 18(3), 205–224, 1993.

[18] B. Baynat und Y. Dallery. A product–form approximation method for general closed queueing networks with several classes of customers. *Performance Evaluation*, 24(3), 165–188, 1996.

[19] B. Baynat und Y. Dallery. An approximation method for general closed queueing networks with Fork/Join mechanism. *Journal of the Operational Research Society*, 51(2), 198–208, 2000.

[20] H. Beilner. Skript zur Vorlesung Warteschlangennetze. Universität Dortmund, WS 1995/96.

[21] H. Beilner, J. Mäter und N. Weißenberg. Towards a Performance Modelling Environment: News on HIT. In *Proc. 4th Int. Conf. on Modelling Techniques and Tools for Computer Performance Evaluation*. 1988.

[22] H. Beilner, J. Mäter und C. Wysocki. The Hierarchical Evaluation Tool HIT. Short Papers and Tool Descriptions of the 7th Int. Conf. on Modelling Techniques and Tools for Computer Performance Evaluation, 1994.

[23] G. Bitrani und D. Tirupati. Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Science*, 34(1), 75–100, 1988.

[24] G. Bolch, S. Greiner, H. de Meer und K. Trivedi. *Queueing Networks and Markov Chains*. Wiley, 1998.

[25] S. Brin und L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1), 107–117, 1998.

[26] J. Buzen. Computational Algorithms for Closed Queueing Networks with Exponential Servers. *Communications of the ACM*, 16(9), 527–531, 1973.

[27] C. Cassandras und S. Lafortune. *Introduction to Discrete Event Systems*. Kluwer, 1999.

[28] K. Chandy, U. Herzog und L. Woo. Parametric Analysis of Queueing Networks. *IBM Journal of Research and Development*, 12(1), 36–42, 1975.

[29] K. Chandy und C. Sauer. Computational Algorithms for Product Form Queueing Networks. *Communications of the ACM*, 23(10), 573–583, 1980.

[30] K.-C. Chang, B. He, C. Li, M. Patel und Z. Zhang. Structured Databases on the Web. Observations and Implications. *SIGMOD Record*, 33(3), 61–70, 2004.

[31] A. Conway und N. Georganas. RECAL – A New Efficient Algorithm for the Exact Analysis of Multiple–Chain Closed Queuing Networks. *Journal of the ACM*, 33(4), 768–791, 1986.

[32] D. Cox. A use of complex probabilities in the theory of stochastic processes. In *Proc. Camb. Phil. Soc.*, 51, 313–319. 1955.

[33] J. Daigle. *Queueing Theory for Telecommunications*. Addison Wesley, 1992.

[34] E. de Souza e Silva und S. Lavenberg. Calculating Joint Queue–Length Distributions in Product-Form Queuing Networks. *Journal of the ACM*, 36(1), 194–207, 1989.

[35] L. Flatto. Two Parallel Queues Created by Arrivals with Two Demands II. *SIAM Journal on Applied Mathematics*, 45(5), 861–878, 1984.

[36] L. Flatto und S. Hahn. Two Parallel Queues Created by Arrivals with Two Demands I. *SIAM Journal on Applied Mathematics*, 44(5), 1041–1053, 1984.

[37] E. Gelenbe und G. Pujolle. *Introduction to Queueing Networks*. Wiley, 1998.

[38] W. Gordon und G. Newell. Closed queueing systems with exponential servers. *Operations Research*, 15(2), 254–265, 1967.

[39] U. Harder und P. Harrison. A Queueing Network Model of Oracle Parallel Server. In *UK Performance Evaluation Workshop*, 69–80. 1999.

[40] B. Haverkort. Approximate Analysis of Networks of PH/PH/1/K Queues: Theory & Tool Support. In *Quantitative Evaluation of Computing and Communication Systems*, LNCS 977, 239–253. Springer, 1995.

[41] B. Haverkort. Approximate Analysis of Networks of PH/PH/1/K Queues with Customer losses: Test results. *Annals of Operations Research*, 79, 271–291, 1998.

[42] A. Heindl und M. Telek. MAP-based decomposition of tandem networks of ·/PH/1/(K) queues with MAP input. In *Proc. 11th GI/ITG Conf. on Measurement, Modelling and Evaluation of Computer and Communication Systems*, 179–194. Aachen, Germany, 2001.

[43] A. Heindl und M. Telek. Output models of MAP/PH/1/(K) queues for an efficient network decomposition. *Performance Evaluation*, 49(1-4), 321–339, 2002.

[44] J. Jackson. Networks of waiting lines. *Operations Research*, 5, 518–521, 1957.

[45] J. Jackson. Jobshop–like queueing systems. *Management Sciences*, 10, 131–142, 1963.

[46] G. Keller, M. Nüttgens und A. Scheer. Semantische Prozeßmodellierung auf der Grundlage „Ereignisgesteuerter Prozeßketten (EPK)“. Veröffentlichungen des Instituts für Wirtschaftsinformatik Saarbrücken, Heft 89, 1992.

[47] C. Kim und A. Agrawala. Analysis of the fork-join queue. *IEEE Transactions on Computers*, 38(2), 250–255, 1989.

[48] E. Kindler. On the semantics of EPCs: A framework for resolving the vicious circle. 2. GI-Workshop: EPK 2003 - Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten, Bamberg, 2003.

[49] L. Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[50] S. Ko und R. Serfozo. Response Times in M/M/s Fork-Join Networks. *Advances in Applied Probability*, 36(3), 854–871, 2004.

[51] R. Koether (Hrsg). *Taschenbuch der Logistik*. Hanser Verlag, 2004.

[52] S. Koukoumialos und G. Liperopoulos. An Analytical Method for the Performance Evaluation of Echelon Kanban Control Systems. erscheint in OR Spektrum, 2005.

[53] W. Krämer und M. Langenbach-Belz. Approximate formulae for the delay in the queueing system GI/G/1. In *Proc. of 8th Int. Teletraffic Congress*, 235/1–8. Melbourne, Australia, 1976.

[54] A. Krishnamurthy und R. Suri. Analytical Models for Pull-type Control Strategies in Manufacturing Systems. In *Proc. of the Industrial Engineering Research Conference*. Houston, 2004.

[55] A. Krishnamurthy, R. Suri und M. Vernon. A New Approach for Analyzing Queueing Models of Material Control Strategies in Manufacturing Systems. In *Proc. 4th Int. Workshop on Queueing Networks with Finite Capacity (QNETs2000)*. West Yorkshire, U.K., 2000.

[56] A. Krishnamurthy, R. Suri und M. Vernon. Two–Moment Approximations for Throughput and Mean Queue Length of a Fork/Join Station with General Input Processes. In J. Shanthikumar, D. Yao und W. Zijm (Hrsg), *Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains*, 87–126. Kluwer International Series in Operations Research and Management Science, 2003.

[57] A. Krishnamurthy, R. Suri und M. Vernon. Analysis of a Fork/Join Station with Inputs from Coxian Servers in a Closed Queueing Network. *Annals of Operations Research*, 25, 69–94, 2004.

[58] A. Kuhn. *Prozeßketten in der Logistik, Entwicklungstrends und Umsetzungsstrategien*. Verlag Praxiswissen, 1995.

[59] P. Kühn. Analysis of complex queueing networks by decomposition. 8-th International Teletraffic Congress, 236-1 - 236-8, Melbourne, 1976.

[60] P. Kühn. Approximate Analysis of General Queueing Networks by Decomposition. *IEEE Transactions on Communications*, 27(1), 1979.

[61] G. Latouche und V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, 1999.

[62] S. Lavenberg. *Computer Performance Modelling Handbook*. Academic Press, 1982.

[63] E. Lazowska, J. Zahorjan, G. Graham und K. Sevcik. *Quantitative System Performance. Computer System Analysis using Queueing Netwok Models*. Prentice-Hall, 1984.

[64] J. Lui, R. Muntz und D. Towsley. Computing Performance Bounds of Fork-Join Parallel Programs under a Multiprocessing Environment. *IEEE Transaction on Parallel and Distributed Systems*, 9(3), 295–311, 1998.

[65] R. Marie. An Approximate Analytical Method for General Queueing Networks. *IEEE Transactions on Software Engineering*, 5(5), 530–538, 1979.

[66] R. Marie. Calculating Equilibrium Probabilities for $\lambda(n)/C_k/1/\mathrm{N}$ Queues. *ACM Sigmetrics Performance Evaluation Review*, 9(2), 117–125, 1980.

[67] K. Marshall. Bounds for some generalisations of the GI/G/1 queue. *Operations Research*, 16, 841–848, 1968.

[68] K. Marshall. Some Inequalities in Queueing. *Operations Research*, 16, 651–665, 1968.

[69] R. Nelson. *Probability, Stochastic Processes and Queueing Theory*. Springer, 1995.

[70] R. Nelson und A. Tantawi. Approximate Analysis of Fork/Join Synchronization in Parallel Queues. *IEEE Transactions on Computers*, 37(6), 739–743, 1988.

[71] R. Nelson, D. Towsley und A. Tantawi. Performance Analysis of Parallel Processing Systems. *IEEE Transactions on Software Engineering*, 14(4), 532–540, 1988.

[72] D. Neuse und K. Chandy. SCAT: A Heuristic Algorithm for Queueing Network Models of Computing Systems. *ACM Sigmetrics Performance Evaluation Review*, 10(2), 59–79, 1981.

[73] M. Neuts. *Matrix–Geometric Solutions in Stochastic Models*. John Hopkins University Press, 1981.

[74] M. Reiser und S. Lavenberg. Mean–Value Analysis of Closed Multichain Queuing Networks. *Journal of the ACM*, 27(2), 313–322, 1980.

[75] R. Sadre und B. Haverkort. Flows in Networks of MAP/MAP/1 Queues. In *Proc. 11th GI/ITG Conf. on Measuring, Modelling and Evaluation of Computer and Communication Systems*, 195–208. Aachen, Germany, 2001.

[76] R. Sadre, B. Haverkort und A. Ost. An Efficient and Accurate Decomposition Method for Open Finite- and Infinite-Buffer Queueing Networks. In *Proc. 3rd Int. Workshop Numerical Solution of Markov Chains*, 1–20. Zaragoza, Spain, 1999.

[77] C. Sauer und K. Chandy. *Computer Systems Performance Modelling*. Prentice-Hall, 1981.

[78] A. Scheer und W. Jost. *ARIS in der Praxis - Gestaltung, Implementierung und Optimierung von Geschäftsprozessen*. Springer, 2002.

[79] P. Schweitzer. Approximate Analysis of Multiclass Closed Networks of Queues. In *Proc. Int. Conf. on Stochastic Control and Optimization*, 25–29. Amsterdam, Netherlands, 1979.

[80] H. Seidlmeier. *Prozessmodellierung mit ARIS*. Vieweg, 2002.

[81] E. Selberg und O. Etzioni. The MetaCrawler Architecture for Resource Aggregation on the Web. *IEEE Expert*, 12(1), 8–14, 1997.

[82] L. Si und J. Callan. Relevant document distribution estimation method for resource selection. In *Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. Toronto, 2003.

[83] D. Stoyan. *Qualitative Eigenschaften und Abschätzungen stochastischer Modelle*. Oldenbourg, 1977.

[84] H. Tijms. *Stochastic models - An algorithmic approach*. Wiley, 1995.

[85] W. van der Aalst, J. Desel und E. Kindler. On the semantics of EPCs: A vicious circle. 1. GI-Workshop: EPK 2002 - Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten, Universität Trier, 2002.

[86] N. van Dijk. *Queueing Networks and Product Forms. A Systems Approach*. Wiley, 1993.

[87] E. Varki. Response time analysis of closed fork-join networks. Technical Report, University of New Hampshire, 1998.

[88] E. Varki. Mean Value Technique for Closed Fork–Join Networks. In *Proc. of the 1999 ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems*, 103–112. Atlanta, US, 1999.

[89] E. Varki und L. Dowdy. Analysis of balanced fork-join systems. In *Proc. of ACM SIGMETRICS Conf. on Measurement and Modeling of Computer Systems*. 1996.

[90] E. Varki und S. Wang. A performance model of disk array storage systems. In *The Computer Measurement Group's 2000 International Conference*. Orlando, US, 2000.

[91] S. Varma und A. Makowski. Interpolation approximations for symmetric fork-join queues. *Performance Evaluation*, 20(1-3), 245–265, 1994.

[92] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, 1988.

[93] W. Whitt. Performance of the Queueing Network Analyzer. *The Bell System Technical Journal*, 62(9), 2817–2843, 1983.

[94] W. Whitt. The Queueing Network Analyzer. *The Bell System Technical Journal*, 62(9), 2779–2815, 1983.

[95] W. Whitt. The Amount of Overtaking in a Network of Queues. *Networks*, 14(3), 411–426, 1984.

[96] W. Whitt. Towards better multi–class parametric–decomposition approximations for open queueing networks. *Annals of Operations Research*, 48, 221–248, 1994.

[97] H. Wildemann. Kundenorientierung und Effizienz in der Automobilindustrie. Symposium: Einsatz von eBusiness-Lösungen in der Automobilindustrie – E-Maturity, Frankfurt a.M., September 2001. www.ematurity.de.

[98] J. Zahorjan und E. Wong. The Solution of Separable Queueing Network Models Using Mean Value Analysis. *ACM Sigmetrics Performance Evaluation Review*, 10(3), 80–85, 1981.

Anhang A

Approximation mit Phasenverteilungen

Ein zentrales Prinzip des in dieser Arbeit angewandten Dekompositionsverfahrens zur Analyse erweiterter Fork/Join–Warteschlangennetze ist die Approximation der Zwischenankunfts– und Bedienzeitverteilungen der betrachteten Stationstypen durch Phasenverteilungen. Diese Vorgehensweise erlaubt die stationäre Analyse der isolierten Stationen anhand ihrer zugrundeliegenden QBD–Struktur. In diesem Anhang werden die betrachteten Phasenverteilungen vorgestellt. Ferner werden einige von Stoyan [83] definierte Halbordnungen auf Verteilungsfunktionen reflektiert und auf die betrachteten Phasenverteilungen angewandt. Die folgende Definition stellt zunächst den Begriff der Phasenverteilung vor.

Definition A.1 *Die Matrix T beschreibe das Verhalten einer endlichen zeitkontinuierlichen homogenen absorbierenden Markovkette in den transienten Zuständen. Die Startverteilung auf den transienten Zuständen sei durch den Vektor τ gegeben. Dann heißt die Zufallsvariable X phasenverteilt, wenn X der Verweilzeit der Markovkette in den transienten Zuständen bis zur Absorption entspricht. $F_X(t) = P\{X \leq t\}$ ist folglich die Verteilung der Absorptionszeit von T unter der Startverteilung τ. Übliche Bezeichnungen sind: X ist $PH(T, \tau)$–verteilt oder X ist phasenverteilt mit der Repräsentation (T, τ).*

Nach Neuts [73] besitzt die Verteilungsfunktion einer durch (T, τ) repräsentierten Phasenverteilung die Darstellung:

$$F_X(t) = 1 - \tau e^{Tt}\mathbf{e}. \tag{A.1}$$

Ferner sind die k–ten Momente $E[X^k]$ der Verteilung von X gegeben durch:

$$E[X^k] = (-1)^k k! \tau T^{-k}\mathbf{e}. \tag{A.2}$$

Insbesondere ergibt sich damit der Erwartungswert $E[X]$ zu:

$$E[X] = -\tau T^{-1}\mathbf{e}. \tag{A.3}$$

In diesem Anhang wird die Anpassung beliebiger kontinuierlicher nicht–negativer Verteilungen mit endlichem ersten und zweiten Moment durch geeignete Phasenverteilungen vorgestellt. Das Ziel liegt genauer darin, unter der Kenntnis des Erwartungswertes $E[Y]$ und des Variationskoeffizienten C_Y der Verteilung einer Zufallsvariablen Y die Parameter der Phasenverteilung einer Zufallsvariablen X derart zu bestimmen, daß gilt:

$$\begin{array}{rcl} E[Y] & = & E[X] \\ C_Y & = & C_X. \end{array} \tag{A.4}$$

Die Wahl des konkreten Typs der Phasenverteilung von X läßt sich ausschließlich aufgrund des Variationskoeffizienten C_Y der Verteilung von Y treffen. In dieser Arbeit wird die in der Tabelle A.1 dargestellte Unterscheidung angenommen. Die Anpassung einer Verteilung mit

C_Y	Phasentyp	Anzahl Phasen
$= 1$	Negativ–Exponentialverteilung	1
> 1	Hyper–Exponentialverteilung mit balanced means	2
< 1	Hypo–Exponentialverteilung	$\left\lceil \frac{1}{C_Y^2} \right\rceil$

Tabelle A.1: Wahl des Typs der Phasenverteilung abhängig von C_Y

$C_Y = 1$ hinsichtlich der ersten beiden Momente durch eine negativ–exponentiell verteilte Zufallsvariable ist offensichtlich, so daß in den folgenden Abschnitten lediglich auf den Fall $C_Y \neq 1$ eingegangen wird.

Bevor jedoch die Anpassung durch Phasenverteilungen vorgestellt wird, faßt der folgende Abschnitt zunächst einige Resultate von Stoyan [83] über Halbordnungen auf Verteilungsfunktionen zusammen. Auf dieser Grundlage werden schließlich Monotonieaussagen über die Wartezeitverteilung in GI/GI/1/∞–Systemen abgeleitet, die essentiell für die Konvergenz des Iterationsverfahrens aus Kapitel 6 sind.

A.1 Halbordnungen auf Verteilungsfunktionen

Dieser Abschnitt reflektiert die von Stoyan [83] definierten Halbordnungen auf Verteilungsfunktionen sowie einige daraus resultierende Folgerungen.

Definition A.2 *Die Verteilungsfunktion F_1 heißt kleiner als F_2 bzgl. $\overset{(1)}{\leq}$ (symbolisch $F_1 \overset{(1)}{\leq} F_2$), wenn für alle $t \in \mathbb{R}$ die Beziehung*

$$F_1(t) \geq F_2(t) \tag{A.5}$$

erfüllt ist. Sind X_1 und X_2 Zufallsvariablen mit den Verteilungsfunktionen F_1 und F_2, so bedeutet $X_1 \overset{(1)}{\leq} X_2$ die Gültigkeit von (A.5).

Definition A.3 *Die Verteilungsfunktion F_1 heißt kleiner als F_2 bzgl. $\overset{(2)}{\leq}$ (symbolisch $F_1 \overset{(2)}{\leq} F_2$), wenn für alle $x \in \mathbb{R}$ die Beziehung*

$$\int_x^\infty (1 - F_1(t))dt \leq \int_x^\infty (1 - F_2(t))dt \tag{A.6}$$

erfüllt ist, wobei

$$\int_x^\infty (1 - F_k(t))dt < \infty; \quad k = 1, 2 \tag{A.7}$$

vorauszusetzen ist. Sind X_1 und X_2 Zufallsvariablen mit den Verteilungsfunktionen F_1 und F_2, so bedeutet $X_1 \overset{(2)}{\leq} X_2$ die Gültigkeit von (A.6).

Definition A.4 *Die Verteilungsfunktion F_1 heißt kleiner als F_2 bzgl. $\overset{(3)}{\leq}$ (symbolisch $F_1 \overset{(3)}{\leq} F_2$), wenn für alle $x \in \mathbb{R}$ die Beziehung*

$$\int_{-\infty}^x F_1(t)dt \geq \int_{-\infty}^x F_2(t)dt \tag{A.8}$$

erfüllt ist, wobei

$$\int_{-\infty}^0 F_k(t)dt < \infty; \quad k = 1, 2. \tag{A.9}$$

vorauszusetzen ist. Sind X_1 und X_2 Zufallsvariablen mit den Verteilungsfunktionen F_1 und F_2, so bedeutet $X_1 \overset{(3)}{\leq} X_2$ die Gültigkeit von (A.8).

In [83] werden ferner die folgenden Beziehungen unter diesen Halbordnungen gezeigt.

Folgerung A.1 *Es gelten:*

1. $F_1 \overset{(1)}{\leq} F_2 \Rightarrow F_1 \overset{(2)}{\leq} F_2$, *falls* $\int_0^\infty (1 - F_k(t))dt < \infty; k = 1, 2.$
2. $F_1 \overset{(1)}{\leq} F_2 \Rightarrow F_1 \overset{(3)}{\leq} F_2$, *falls* $\int_{-\infty}^0 F_k(t)dt < \infty; k = 1, 2.$
3. $E[X_1] = E[X_2] \Rightarrow F_1 \overset{(2)}{\leq} F_2 \Leftrightarrow F_2 \overset{(3)}{\leq} F_1.$

4. *Die Relationen* $\overset{(1)}{\leq}$, $\overset{(2)}{\leq}$ *und* $\overset{(3)}{\leq}$ *bleiben unter der Faltungsoperation erhalten, d.h.*

$$F_1 \overset{(i)}{\leq} F_2 \Rightarrow F_1 * F_3 \overset{(i)}{\leq} F_2 * F_3, \quad i = 1, 2, 3.$$

5. *Im Fall gleicher Erwartungswerte und* $F_1(0) = F_2(0) = 0$ *ist* $F_1 \overset{(2)}{\leq} F_2$ *äquivalent zu*

$$\int_0^x (1 - F_1(t))dt \geq \int_0^x (1 - F_2(t))dt \tag{A.10}$$

bzw.

$$\int_0^x F_1(t)dt \leq \int_0^x F_2(t)dt. \tag{A.11}$$

Von zentraler Bedeutung ist das folgende Lemma, das die Momente zweier Verteilungsfunktionen vergleicht, für die eine der Relationen $\overset{(1)}{\leq}$ oder $\overset{(2)}{\leq}$ erfüllt ist.

Lemma A.1 *Für nicht–negative Zufallsvariablen* X_1 *und* X_2 *mit* $X_1 \overset{(i)}{\leq} X_2$ $(i = 1, 2)$ *gilt*

$$E[X_1^k] \leq E[X_2^k], \quad \forall k \geq 1, \tag{A.12}$$

sofern die Momente existieren.

A.2 Monotonieeigenschaften für GI/GI/1–FCFS Systeme

In diesem Abschnitt werden einige Aussagen aus [83] vorgestellt, die die Bedeutung der Halbordnungen für die Wartezeitverteilung in GI/GI/1/∞–Systemen verdeutlichen. Diese Aussagen haben essentielle Auswirkungen hinsichtlich der Aggregat–Bestimmung in Kapitel 6.

Satz A.1 *Seien* G_1 *und* G_2 *GI/GI/1/*∞*-FCFS Systeme mit den Zwischenankunftszeitverteilungen* A_1 *und* A_2 *sowie den Bedienzeitverteilungen* B_1 *und* B_2*.*
Dann gilt für die stationären Wartezeitverteilungen W_1 *und* W_2 *die folgende Beziehung:*

$$A_2 \overset{(3)}{\leq} A_1 \text{ und } B_1 \overset{(2)}{\leq} B_2 \Rightarrow W_1 \overset{(2)}{\leq} W_2, \tag{A.13}$$

falls die Erwartungswerte der Durchlaufzeitverteilungen endlich sind.

Satz A.2 *Seien die Voraussetzungen des Satzes A.1 erfüllt. Dann gilt für die Semiinvarianten n–ter Ordnung* $s_{n,1}$ *und* $s_{n,2}$ *der stationären Wartezeiten der beiden GI/GI/1/*∞*–FCFS Systeme für alle* $n \geq 1$ *die Beziehung:*

$$s_{n,1} \leq s_{n,2}. \tag{A.14}$$

Aus den Sätzen A.1 und A.2 ergibt sich schließlich:

Folgerung A.2 *Seien G_1 und G_2 GI/GI/1/∞–FCFS Systeme mit identischen Zwischenankunftszeitverteilungen A und den Bedienzeitverteilungen B_1 und B_2 mit $B_1 \overset{(i)}{\leq} B_2$ (i = 1 oder 2), $E[B_1] = E[B_2]$ und $\sigma_B = Var(B_1) < \sigma_B = Var(B_2)$. Dann sind alle Semiinvarianten der stationären Wartezeit von G_1 strikt kleiner als die von G_2.*

Beweis:

Aus dem Beweis des Satzes A.2 in [83] geht hervor, daß unter obigen Voraussetzungen entweder die Semiinvarianten n-ter Ordnung $s_{n,1}$ der stationären Wartezeit von G_1 für alle $n \geq 1$ strikt kleiner sind als die von G_2 oder aber für alle $n \geq 1$ mit denen von G_2 übereinstimmen. Im ersten Fall ist die Aussage des Satzes bewiesen. Im zweiten Fall sind offensichtlich die Verteilungen der stationären Wartezeit beider Systeme identisch. Aufgrund gleicher Ankunftsverteilungen sind folglich auch die Verteilungen der Pausenzeiten, d.h. der Untätigkeitszeiten der Bediener zwischen einem Bedienende und dem Beginn der nächsten Bedienung, identisch. Nach Marshall [68, 67] läßt sich der Erwartungswert der Wartezeit folgendermaßen darstellen:

$$E[W] = \frac{(E[A] - E[B])^2 + \sigma_A^2 + \sigma_B^2}{2(E[A] - E[B])} - \frac{E[L]^2 + \sigma_L^2}{2E[L]}. \tag{A.15}$$

Dabei sind $E[L]$ und σ_L^2 der Erwartungswert und die Varianz der Pausenzeiten. Aus Gleichung (A.15) folgt unmittelbar, daß die Varianzen der Bedienzeiten B_1 und B_2 identisch sind. Dies ist ein Widerspruch zur Voraussetzung.

□

A.3 $\mathcal{H}^{1+}$–Verteilungen

In diesem Abschnitt wird die $\mathcal{H}^{1+}$–Verteilung auf Basis einer Hyper–Exponentialverteilung mit zwei Phasen definiert. Eine Zufallsvariable X ist gemäß einer zweiphasigen Hyper–Exponentialverteilung verteilt, wenn X mit der Wahrscheinlichkeit p $(0 < p < 1)$ die Werte einer ersten negativ–exponentiell verteilten Zufallsvariable X_1 und mit der alternativen Wahrscheinlichkeit $1 - p$ die einer zweiten negativ-exponentiell verteilten Zufallsvariable X_2 annimmt. Die Abbildung A.1 stellt diese Konstruktion graphisch dar. Es ist leicht einzusehen,

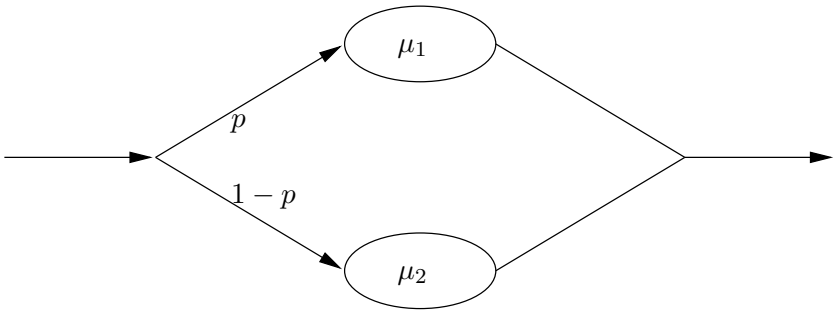

Abbildung A.1: Graphische Darstellung der Hyper-Exponentialverteilung

daß X eine Phasenverteilung mit der Repräsentation (A, α) besitzt mit:

$$\begin{aligned} \alpha &= (p, 1-p)^T \\ A &= \begin{pmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{pmatrix}. \end{aligned} \tag{A.16}$$

Folglich besitzt die Verteilung von X die Darstellung:

$$F_X(t) = 1 - \alpha e^{At}\mathbf{e} = 1 - pe^{-\mu_1 t} - (1-p)e^{-\mu_2 t}. \tag{A.17}$$

Daraus ergeben sich die ersten beiden Momente zu:

$$\begin{aligned} E[X] &= p\tfrac{1}{\mu_1} + (1-p)\tfrac{1}{\mu_2} \\ E[X^2] &= p\tfrac{2}{\mu_1^2} + (1-p)\tfrac{2}{\mu_2^2}. \end{aligned} \tag{A.18}$$

In diesem Abschnitt wird gezeigt, daß dieser Verteilungstyp geeignet ist, beliebige reellwertige nicht–negative Zufallsvariablen Y mit endlichem Erwartungswert $E[Y] > 0$ und endlichem Variationskoeffizienten $C_Y > 1$ hinsichtlich der ersten beiden Momente exakt anzupassen.

Seien also $E[Y]$ und C_Y bekannt. Die Aufgabe besteht folglich darin, die Parameter μ_1, μ_2 und p einer zweiphasig hyper–exponentiell verteilten Zufallsvariablen X derart zu bestimmen, daß gilt:

$$E[X] = E[Y] \tag{A.19}$$

$$C_X = C_Y. \tag{A.20}$$

Der Freiheitsgrad, der aus der Anpassung lediglich zweier Momente durch die drei Parameter μ_1, μ_2 und p resultiert, wird durch die sog. balanced means folgendermaßen ausgefüllt:

$$\frac{p}{\mu_1} = \frac{1-p}{\mu_2} \quad \text{bzw.} \quad \mu_2 = \frac{1-p}{p}\mu_1. \tag{A.21}$$

Damit hat die Verteilungsfunktion F_X von X die Darstellung:

$$F_X(t) = 1 - pe^{-\mu_1 t} - (1-p)e^{-\frac{1-p}{p}\mu_1 t}, \tag{A.22}$$

und die ersten beiden Momente vereinfachen sich zu:

$$\begin{aligned} E[X] &= \tfrac{2p}{\mu_1} \\ E[X^2] &= \tfrac{2p}{(1-p)\mu_1^2}. \end{aligned} \tag{A.23}$$

Ferner besitzt der Variationskoeffizient von X die Form:

$$C_X^2 = 1 + \frac{(2p-1)^2}{2p(1-p)}. \tag{A.24}$$

Offensichtlich hängt also der Variationskoeffizient C_X ausschließlich von dem Parameter p ab. Die Auflösung der Gleichung (A.24) nach p und Einsetzen der Identität (A.20) liefert für p den Wert:

$$p = \frac{1}{2}\left(1 + \sqrt{\frac{C_Y^2 - 1}{C_Y^2 + 1}}\right). \tag{A.25}$$

Es ist leicht einzusehen, daß p als Funktion in C_Y auf dem Bereich $]1, \infty[$ wohldefiniert ist und die Funktionswerte (streng monoton steigend) in den Bereich $]\frac{1}{2}, 1[$ abbildet. Zu einem gegebenen C_Y ist der Parameter $p > \frac{1}{2}$ der hyper–exponentiell verteilten Zufallsvariable X mit balanced means somit eindeutig bestimmt. Für die Parameter μ_1 und μ_2 ergeben sich schließlich die ebenfalls eindeutigen Darstellungen:

$$\mu_1 = \frac{2p}{E[Y]} \tag{A.26}$$

$$\mu_2 = \frac{2(1-p)}{E[Y]}. \tag{A.27}$$

Die somit gefundenen Resultate werden in dem folgenden Satz zusammengefaßt.

Satz A.3 *Sei Y eine beliebige kontinuierliche nicht–negative Zufallsvariable mit dem endlichen Erwartungswert $E[Y] > 0$ und dem endlichen Variationskoeffizienten $C_Y > 1$. Dann existiert eine hyper–exponentiell verteilte Zufallsvariable X mit zwei Phasen und balanced means, deren Verteilungsparameter μ_1, μ_2 und p durch (A.26), (A.27) und (A.25) eindeutig bestimmt sind, und deren Momente die Eigenschaften (A.19) und (A.20) besitzen.*
X heißt im folgenden $\mathcal{H}^{1+}$–verteilt.

Im folgenden werden einige Eigenschaften der $\mathcal{H}^{1+}$–Verteilung zusammengetragen. Zunächst hat die Laplace–Transformierte $L_X(s)$ einer $\mathcal{H}^{1+}$ verteilten Zufallsvariable X mit dem Erwartungswert E und dem Variationskoeffizienten C die Form:

$$L_X(s) = \frac{\frac{2}{E}(sC^2 + \frac{1}{E})}{s^2(C^2+1) + \frac{2s}{E}(C^2+1) + \frac{2}{E^2}}. \tag{A.28}$$

Weiterhin lassen sich $\mathcal{H}^{1+}$ Verteilungen hinsichtlich der in Abschnitt A.1 definierten Halbordnungen in Beziehung setzen, wie die folgenden Sätze zeigen.

Satz A.4 *Seien X_1 bzw. X_2 $\mathcal{H}^{1+}$–verteilt mit den Erwartungswerten $0 < E_1 < E_2 < \infty$ und identischem Variationskoeffizient $1 < C < \infty$. Dann gilt für die Verteilungsfunktionen F_1 bzw. F_2 von X_1 bzw. X_2:*

$$F_1 \overset{(1)}{\leq} F_2. \tag{A.29}$$

<u>*Beweis:*</u>

Seien μ_1, μ_2, p und $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{p}$ die Parameter der Verteilungen F_1 und F_2. Dann gelten aufgrund der Gleichungen (A.25), (A.26) und (A.27) die folgenden Beziehungen:

$$\begin{aligned} \tilde{p} &= p \\ \tilde{\mu}_1 &= l\mu_1 \\ \tilde{\mu}_2 &= l\mu_2 \end{aligned}$$

mit $0 < l = \frac{E_1}{E_2} < 1$. Ist (A, α) die Phasenrepräsentation von F_1 gemäß (A.16), so ist (lA, α) die entsprechende Darstellung von F_2. Wegen $0 < l < 1$ und aufgrund der Monotonie von F_1 folgt:

$$F_1(t) \geq F_1(lt) = 1 - \alpha e^{lAt} = F_2(t). \tag{A.30}$$

Mit der Definition A.2 folgt die Behauptung.

□

Eine zu Satz A.4 analoge Beziehung läßt sich ebenfalls bzgl. der Variationskoeffizienten herleiten. Dabei ist jedoch die Relation $\overset{(1)}{\leq}$ durch $\overset{(2)}{\leq}$ zu ersetzen. Konkret gilt folgender Satz:

Satz A.5 *Seien X_1 bzw. X_2 $\mathcal{H}^{1+}$–verteilt mit identischem Erwartungswert $0 < E < \infty$ und den Variationskoeffizienten $1 < C_1 < C_2 < \infty$. Dann gilt für die Verteilungsfunktionen F_1 bzw. F_2 von X_1 bzw. X_2:*

$$F_1 \overset{(2)}{\leq} F_2. \tag{A.31}$$

Beweis:

Nach Definition A.3 und Folgerung A.1 ist zu zeigen:

$$\int_0^x (1 - F_1(t))dt \geq \int_0^x (1 - F_2(t))dt, \quad \forall x \geq 0. \tag{A.32}$$

Im folgenden wird allgemeiner gezeigt, daß obiges Integral für $\mathcal{H}^{1+}$–Verteilungen mit identischen Erwartungswerten für beliebige $x \geq 0$ in C monoton fällt. Sei also eine derartige Verteilung F betrachtet. Dann besitzt F durch Einsetzen von (A.26) und (A.27) die Darstellung

$$F(t) = 1 - pe^{-\frac{2pt}{E}} - (1-p)e^{-\frac{2(1-p)t}{E}}. \tag{A.33}$$

Ferner gilt:

$$\int_0^x (1 - F(t))dt = E - \frac{E}{2}e^{-\frac{2px}{E}} - \frac{E}{2}e^{-\frac{2(1-p)x}{E}}. \tag{A.34}$$

Da p streng monoton in C steigt (vgl. (A.25)), reicht es aus zu zeigen, daß (A.34) streng monoton in p fällt. Die Ableitung $h(p)$ der rechten Seite von (A.34) nach p hat die Form:

$$h(p) = x\left(e^{-\frac{2px}{E}} - e^{-\frac{2(1-p)x}{E}}\right). \tag{A.35}$$

Da mit $1 > p > \frac{1}{2}$ auch $p > 1 - p$ gilt, folgt für $x \geq 0$ und $E > 0$ unmittelbar $h(p) \leq 0$ und damit die Behauptung.

□

A.4 $\mathcal{H}^{1-}$–Verteilungen

In diesem Abschnitt wird der Begriff einer $\mathcal{H}^{1-}$–Verteilung auf der Grundlage einer Hypo–Exponentialverteilung mit K Phasen definiert. Eine Hypo–Exponentialverteilung mit $K \geq 2$ Phasen setzt sich aus K aufeinanderfolgenden negativ–exponentiell verteilten Phasen mit den Parametern $\mu_1, \ldots, \mu_K$ zusammen. Diese Konstruktion ist in der Abbildung A.2 graphisch dargestellt.

Abbildung A.2: Graphische Darstellung der Hypo-Exponentialverteilung mit K Phasen

Dieser Verteilungstyp besitzt die Phasenrepräsentation (A, α) mit:

$$\alpha = (1, 0, \ldots, 0)^T$$

$$A = \begin{pmatrix} -\mu_1 & \mu_1 & & & \\ & -\mu_2 & \mu_2 & & \\ & & \ddots & \ddots & \\ & & & -\mu_{K-1} & \mu_{K-1} \\ & & & & -\mu_K \end{pmatrix}. \tag{A.36}$$

Der Erwartungswert $E[X]$ und die Varianz $V[X]$ einer Zufallsvariablen X mit diesem Verteilungstyp ergeben sich durch:

$$E[X] = \sum_{i=1}^{K} \frac{1}{\mu_i} \tag{A.37}$$

$$V[X] = \sum_{i=1}^{K} \frac{1}{\mu_i^2}. \tag{A.38}$$

Da offensichtlich stets $V[X] \leq E[X]^2$ gilt, folgt für den Variationskoeffizienten $C_X \leq 1$.

Dieser Abschnitt verdeutlicht, daß sich dieser Verteilungstyp dazu eignet, beliebige kontinuierliche nicht–negative Zufallsvariablen Y mit positivem Erwartungswert $E[Y]$ und Variationskoeffizienten $0 < C_Y < 1$ exakt hinsichtlich der ersten beiden Momente anzupassen.

Sei also wiederum von der Kenntnis des Erwartungswertes $E[Y]$ und des Variationskoeffizienten C_Y von Y ausgegangen. Das Ziel besteht in der Bestimmung der Parameter $\mu_1, \ldots, \mu_K$ einer K–phasigen Hypo–Exponentialverteilung der Zufallsvariablen X derart, daß gilt:

$$E[X] = E[Y] \tag{A.39}$$

$$C_X = C_Y. \tag{A.40}$$

Dazu sei zunächst davon ausgegangen, die Anzahl $K \geq 2$ der Phasen sei bekannt. Dann existieren K Freiheitsgrade zur Anpassung zweier Momente. Es wird sich zeigen, daß die Wahl $\mu_2 = \cdots = \mu_K$ geeignet ist, so daß lediglich zwei Freiheitsgrade übrig bleiben.

Damit vereinfachen sich der Erwartungswert und die Varianz von X zu:

$$E[X] = \frac{1}{\mu_1} + \frac{K-1}{\mu_2} \tag{A.41}$$

$$V[X] = \frac{1}{\mu_1^2} + \frac{K-1}{\mu_2^2}. \tag{A.42}$$

Ferner besitzt der Variationskoeffizient C_X die Darstellung:

$$C_X^2 = \frac{\mu_1^2(K-1) + \mu_2^2}{(\mu_1(K-1) + \mu_2)^2}. \tag{A.43}$$

Aus den Gleichungen (A.41) und (A.43) lassen sich die Parameter μ_1 und μ_2 der angepaßten Hypo–Exponentialverteilung angeben zu:

$$\mu_1 = \frac{\sqrt{C_Y^2K^2 - C_Y^2K - K + 1)} - 1}{E[Y](C_Y^2K - C_Y^2 - 1)} \leq \frac{K}{E[Y]} \tag{A.44}$$

$$\mu_2 = \frac{(K-1)(\sqrt{C_Y^2K^2 - C_Y^2K - K + 1)} - 1)}{E[Y](\sqrt{C_Y^2K^2 - C_Y^2K - K + 1)} - C_Y^2K + C_Y^2)} \geq \frac{K}{E[Y]}. \tag{A.45}$$

Leicht ergibt sich für die Existenz der μ_1 und μ_2, die die Gleichungen (A.39) und (A.40) erfüllen, die Bedingung:

$$C_Y^2 \geq \frac{1}{K}. \tag{A.46}$$

Mit K Phasen lassen sich folglich Verteilungen mit Variationskoeffizienten im Bereich $\frac{1}{K} \leq C_Y^2 < 1$ approximieren. Eine geeignete Wahl der Phasenanzahl K ist somit:

$$K = \left\lceil \frac{1}{C_Y^2} \right\rceil. \tag{A.47}$$

Im Fall $K = 1/C_Y^2$ ergibt sich $\mu_1 = \mu_2 = K/E[Y]$. Ferner ist μ_1 eine in C_Y streng monoton fallende und μ_2 eine in C_Y streng monoton steigende Funktion. Für $C_Y^2 \to 1$ konvergiert μ_2 gegen unendlich und μ_1 gegen $1/E[Y]$.

Diese Resultate werden in folgendem Satz zusammengefaßt.

Satz A.6 *Sei Y eine beliebige nicht–negative Zufallsvariable mit dem endlichen Erwartungswert $E[Y]$ und dem Variationskoeffizienten C_Y mit $0 < C_Y < 1$. Dann existiert eine hypo–exponentiell verteilte Zufallsvariable X mit $K \geq 1/C_Y^2$ Phasen und den Raten $\mu_1, \mu_2 = \cdots = \mu_K$ derart, daß die ersten beiden Momente von X und Y übereinstimmen. Die Werte μ_1, μ_2 sind gemäß Gleichungen (A.44) und (A.45) bestimmt. X heißt im folgenden $\mathcal{H}^{1-}$–verteilt mit den Parametern $K, \mu 1, \mu 2$.*

Abschließend werden einige Eigenschaften von $\mathcal{H}^{1-}$–Verteilungen zusammengefaßt. Zunächst besitzt die Laplace–Transformierte die Darstellung:

$$\begin{aligned} L(s) &= \frac{\left(\frac{(A-1)(K-1)}{B}\right)^K B}{(K-1)(-1+A+sE[Y](C_Y^2K-C_Y^2-1))} \\ A &= \sqrt{(C_Y^2K-1)(K-1)} \\ B &= (sA-s(K-1)C_Y^2)E[Y]+(A-1)(K-1). \end{aligned} \tag{A.48}$$

Weiterhin gelten für $\mathcal{H}^{1-}$–Verteilungen dieselben Aussagen hinsichtlich der in Abschnitt A.1 definierten Halbordnungen, wie sie für die $\mathcal{H}^{1+}$–Verteilungen in den Sätzen A.4 und A.5 bereits bewiesen wurden. Diese Tatsache wird im folgenden bewiesen.

Satz A.7 *Seien X_1 bzw. X_2 $\mathcal{H}^{1-}$–verteilt mit den Erwartungswerten mit $0 < E_1 < E_2 < \infty$ und identischem Variationskoeffizient $0 < C < 1$. Die Phasenanzahl sei ebenfalls identisch. Dann gilt für die zugehörigen Verteilungsfunktionen F_1 und F_2:*

$$F_1 \overset{(1)}{\leq} F_2. \tag{A.49}$$

<u>*Beweis:*</u>

Aufgrund der Gleichungen (A.44) und (A.45) unterscheiden sich die Parameter μ_1, μ_2 von F_1 und $\tilde{\mu}_1, \tilde{\mu}_2$ von F_2 um denselben Faktor $l = \frac{E_1}{E_2} < 1$, d.h.

$$\tilde{\mu}_i = l\mu_i, \quad i = 1, 2. \tag{A.50}$$

Ist also (A, α) die Phasenrepräsentation von F_1, so ist (lA, α) die Phasenrepräsentation von F_2. Aus $l < 1$ und der Monotonie von Verteilungsfunktionen folgt:

$$F_1(t) \geq F_1(lt) = 1 - \alpha e^{lAt} = F_2(t). \tag{A.51}$$

Aus der Definition A.2 resultiert schließlich die Behauptung.

□

Im Falle identischer Erwartungswerte gilt der folgende Satz.

Satz A.8 *Seien X_1 bzw. X_2 $\mathcal{H}^{1-}$–verteilt mit dem identischen Erwartungswert $0 < E < \infty$ und den Variationskoeffizienten $0 < C_1 < C_2 < 1$. Die Anzahl K der Phasen beider Verteilung sei identisch, d.h. $K \geq \frac{1}{C_1^2}$. Dann gilt für die zugehörigen Verteilungsfunktionen F_1 und F_2 die Beziehung:*

$$F_1 \overset{(2)}{\leq} F_2. \tag{A.52}$$

Beweis:

Die Wahl der identischen Phasenanzahl K der Verteilungen F_1 und F_2 ist für den Satz wesentlich, da sich die Aussage des Satzes im Falle unterschiedlicher Phasenanzahlen im allgemeinen nicht beweisen läßt. In diesem Fall sind zwei Hypo–Exponentialverteilungen im allgemeinen nicht bzgl. der $\overset{(2)}{\leq}$–Relation vergleichbar.

1. *Der Beweis wird zunächst für $K = 2$ Phasen geführt. Nach Folgerung A.1 ist die folgende Beziehung zwischen den Funktionen F_1 und F_2 zu zeigen:*

$$\int_0^x (1 - F_1(t))dt \geq \int_0^x (1 - F_2(t))dt \quad \forall x \geq 0. \tag{A.53}$$

Wiederum wird allgemeiner gezeigt, daß für $\mathcal{H}^{1-}$–Verteilungen mit den durch (A.44) und (A.45) gegebenen Parametern das Integral $\int_0^x (1 - F(t))dt$ in dem Variationkoeffizienten streng monoton fällt. Dies ist äquivalent dazu, daß das Integral in μ_2 streng monoton fällt bzw. daß die Ableitung $g(\mu_2)$ des obigen Integrals kleiner oder gleich 0 ist. $g(\mu_2)$ hat die Darstellung:

$$g(\mu_2) = \Big(-\Big(\mu_2 x(\mu_2 E - 2) + 2(E\mu_2 - 1)\Big) e^{\frac{-\mu_2 x(-2+E\mu_2)}{(E\mu_2-1)}} \tag{A.54}$$

$$-\Big(\mu_2 x(E\mu_2 - 2) - 2(E\mu_2 - 1)\Big)\Big) \frac{e^{-\frac{\mu_2 x}{(E\mu_2-1)}}}{(E\mu_2 - 2)^2 \mu_2^2}. \tag{A.55}$$

Aus (A.45) folgt $E\mu_2 \geq 2$. Damit ist $g(\mu_2) \leq 0$ äquivalent zu:

$$-\left(\mu_2 x(\mu_2 E - 2) + 2(E\mu_2 - 1)\right) e^{\frac{-\mu_2 x(-2+E\mu_2)}{(E\mu_2-1)}} - \left(\mu_2 x(E\mu_2 - 2) - 2(E\mu_2 - 1)\right) \leq 0. \tag{A.56}$$

Mit der Definition

$$A = \frac{\mu_2 x(E\mu_2 - 2)}{(E\mu_2 - 1)} \geq 0 \quad \forall x \geq 0$$

hat Gleichung A.56 die Darstellung:

$$-(A + 2)e^{-A} - (A - 2) \leq 0 \quad \forall A \geq 0. \tag{A.57}$$

Die linke Seite obiger Gleichung besitzt für $A = 0$ den Wert 0. Die Ableitung der linken Seite nach A nimmt für $A = 0$ ebenfalls den Wert 0 an. Die zweite Ableitung ist für alle $A \geq 0$ stets kleiner oder gleich 0, so daß Gleichung A.57 für alle $A \geq 0$ erfüllt ist.

Zusammenfassend gilt im Fall $K = 2$ $F_1 \overset{(2)}{\leq} F_2$ genau dann, wenn $C_1 \leq C_2$ bzw. wenn für die Parameter μ_1, μ_2 bzw. $\tilde{\mu}_1, \tilde{\mu}_2$ der Funktion F_1 bzw. der Funktion F_2 gilt:

$$\tilde{\mu}_1 \leq \mu_1 \leq \mu_2 \leq \tilde{\mu}_2. \tag{A.58}$$

Damit ist der Beweis im Fall $K = 2$ vollständig.

2. *Der Beweis für eine beliebige Anzahl an Phasen erfolgt durch vollständige Induktion über die Anzahl K der Phasen. Dazu sei die Aussage des Falls $K = 2$ zunächst bzgl. der Parameter der beiden Phasen formuliert. Mit den Gleichungen (A.44) und (A.45) gilt für zweiphasige $\mathcal{H}^{1-}$–Verteilungen F bzw. $\tilde{F}$ mit identischen Erwartungswerten und den Parametern μ_1, μ_2 bzw. $\tilde{\mu}_1, \tilde{\mu}_2$ die Beziehung $F \overset{(2)}{\leq} \tilde{F}$ genau dann, wenn gilt:*

$$\tilde{\mu}_2 \geq \mu_2 \geq \mu_1 \geq \tilde{\mu}_1. \tag{A.59}$$

Dementsprechend läßt sich die Induktionsvoraussetzung folgendermaßen formulieren: Für K–phasige $\mathcal{H}^{1-}$-Verteilungen F bzw. $\tilde{F}$ mit identischen Erwartungswerten, die jeweils eine Phase mit dem Parameter μ_1 bzw. $\tilde{\mu}_1$ und $K-1$ Phasen mit dem Parameter μ_2 bzw. $\tilde{\mu}_2$ besitzen, gilt $F \overset{(2)}{\leq} \tilde{F}$ genau dann, wenn Gleichung A.59 erfüllt ist.

Für den Induktionsschritt seien also $K+1$–phasige $\mathcal{H}^{1-}$–Verteilungen F bzw. $\tilde{F}$ mit identischem Erwartungswerten und jeweils einer Phase mit dem Parameter μ_1 bzw. $\tilde{\mu}_1$ und K Phasen mit dem Parameter μ_2 bzw. $\tilde{\mu}_2$ betrachtet. Ferner besitze die $(K+1)$–phasige Hypo–Exponentialverteilung F' eine Phase mit dem Parameter μ_1, eine Phase mit dem Parameter μ_2' und $K-1$ Phasen mit dem Parameter $\tilde{\mu}_2$. F' besitze zudem denselben Erwartungswert wie F bzw. $\tilde{F}$. Dann folgt unmittelbar $\tilde{\mu}_2 \geq \mu_2 \geq \mu_2'$. Mit der Induktionsvoraussetzung und Folgerung A.1 gilt $F \overset{(2)}{\leq} F'$. Da ebenso $\tilde{\mu}_2 \geq \mu_2', \mu_1 \geq \tilde{\mu}_1$ gilt, folgt $F' \overset{(2)}{\leq} \tilde{F}$ und insgesamt $F \overset{(2)}{\leq} \tilde{F}$.

□

A.5 $\mathcal{H}$–Verteilungen

In diesem Abschnitt werden die Resultate der Abschnitte A.3 und A.4 in den folgenden Definitionen bzw. Sätzen zusammengefaßt.

Definition A.5 *Eine Zufallsvariable X mit dem Erwartungswert $E[X]$ und dem Variationskoeffizienten C_X heißt $\mathcal{H}$–verteilt, wenn sie in den Fällen $C_X < 1$, $C_X = 1$ bzw. $C_X > 1$ $\mathcal{H}^{1-}$, negativ–exponentiell bzw. $\mathcal{H}^{1+}$–verteilt ist.*

Satz A.9 *Für zwei $\mathcal{H}$–Verteilungen F_1 und F_2 mit den Erwartungswerten $0 < E_1 < E_2 < \infty$ und identischem Variationskoeffizient C gilt:*

$$F_1 \overset{(1)}{\leq} F_2.$$

Im Fall $0 < C < 1$ wird zudem gefordert, daß F_1 und F_2 identische Phasenanzahlen besitzen.

Beweis:

Im Fall $C \neq 1$ sind die Beweise in den Sätzen A.4 und A.7 vorgestellt. Der Beweis des Falls $C = 1$ wird in [83] vorgestellt.

□

Satz A.10 *Für zwei $\mathcal{H}$–Verteilungen F_1 und F_2 mit identischem Erwartungswert E und den Variationskoeffizienten $0 < C_1 < C_2 < \infty$ gilt:*

$$F_1 \overset{(2)}{\leq} F_2.$$

Im Fall $0 < C_1 < C_2 < 1$ wird zudem gefordert, daß F_1 und F_2 durch eine identische Anzahl an Phasen repräsentiert sind.

Beweis:

In den Fällen $0 < C_1 < C_2 < 1$ bzw. $1 < C_1 < C_2$ sind die Beweise in den Sätzen A.5 und A.8 vorgestellt. Im Fall $0 < C_1 \leq 1 \leq C_2$ ist leicht einzusehen, daß die negative Exponentialverteilung mit dem Parameter $\frac{1}{E}$ bzgl. der Relation $\overset{(2)}{\leq}$ eine obere bzw. untere Schranke der $\mathcal{H}^{1-}$ bzw. der $\mathcal{H}^{1+}$–Verteilung mit dem Erwartungswert E bildet. Diese Tatsache wird ferner in [83] bewiesen.

□