Technische Universität Carolo-Wilhelmina zu Braunschweig
Institut für Betriebssysteme und Rechnerverbund
Prof. Dr. M. Zitterbart

Handover Priorisierung für Anwendungen mit variablen Bitraten in mobilen DiffServ Netzen

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science

Aufgabenstellung und Betreuer:
Prof. Dr. M. Zitterbart und Dipl.-Inform. Jörg Diederich

Braunschweig, 1.8.2002

Cand. Inform. Carsten Buschmann
Kleine Kreuzstr. 9
38118 Braunschweig

Eigenständigkeitserklärung

Ich versichere, die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Hilfsmittel angefertigt zu haben.

Braunschweig, den 1.8.2002

Carsten Buschmann

Kurzfassung

Die Vermeidung von Verbindungsabbrüchen, die durch Zellenwechsel (Handovers) bedingt sind, ist ein wichtiger Punkt von Konzepten zur Zusicherung von Dienstgüte in zellulären Mobilfunknetzwerken. In dieser Arbeit wird ein Verfahren zur Handoverpriorisierung vorgestellt. Es reduziert die so genannten Handover Drops, indem in den Funkzellen ein bestimmter Betrag der Bandbreitenressourcen speziell für Handovers reserviert wird. Dabei liegt ein besonderes Augenmerk auf Skalierbarkeit und auf Anwendungen mit variabler Bitrate. Die möglichst exakte Ermittlung der von den Datenströmen insgesamt belegten mittleren Bandbreite bildet den Ausgangspunkt für die Festlegung der Reservierungshöhen sowie für die Zugangskontrolle. Dazu kommen aggregierte Messungen der Netzwerkauslastung zum Einsatz. Neben den so erzielten Eigenschaften Effizienz und Skalierbarkeit sind Einfachheit, Robustheit, Fairness und Echtzeitberechenbarkeit weitere Charakteristika des Verfahrens.

Abstract

In order to provide quality of service assurances in cellular networks, avoiding handover drops is an important issue. A scheme for handover priorization is introduced in this paper. It reduces the number of handover drops by reserving a certain amount of bandwidth for handovers in every base station. The design focuses on scalability and applications with variable bitrates. Reservations and admission control are based on a precise estimate of the aggregated average bandwidth of all active data flows. The estimates rely on aggregate measurements of the network traffic. Besides its properties efficiency and scalability, the propoesed concept ist simple, robust, fair and realtime-computable.

Inhalt

Abbildungsverzeichnis

Abkürzungsverzeichnis

AF	Assured Forwarding
BA	Behaviour Aggregate
CBR	Constant Bitrate
DS	Differentiated Services
DSCP	Differentiated Services Code Point
EWMA	Exponential Weighted Moving Average (Messverfahren)
EXPOO	Exponential On Off (Datenquelle)
God	General Operations Director
GoS	Grade of Service
HOD	Handover Drop
IP	Internet Protocol
MBAC	Measurement-Based Admission Control (Messungsbasierte Zugangskontrolle)
MF	Multi-field (Classificaton)
MTU	Maximum Transfer Unit (maximale Paketgröße)
NC	New Call
NCB	New Call Block
ns-2	The Network Simulator 2
PDA	Personal Digital Assistant
PHB	Per-Hop Behaviour
POO	Pareto On Off (Datenquelle)
QoS	Quality of Service (Dienstgüte)
UCL	Upper Control Limit
UMTS	Universal Mobile Telecomunications System
VBR	Variable Bitrate

Verzeichnis der Abkürzungen in Formeln

B	Gesamtbandbreite einer Basisstation
D	Mittlere Datenrate einer Verbindung
F	noch verfügbare (freie) Bandbreite einer Basisstation
F_i	freie Bandbreite der Nachbarstation i
G_{oben}	obere Grenze des Steuerkorridors des Moduls zur Messung der Zellauslastung
G_{unten}	untere Grenze des Steuerkorridors des Moduls zur Messung der Zellauslastung
H	Steuerparameter von HoPVarB: Reservierungshöhe
M	einzelner Messwert des Moduls zur Messung der Zellauslastung
MR	Moving Range; Maß für die Varianz der Messwerte bei Flip-Flop-Verfahren und HoPVarB
N	Gesamtanzahl der Verbindungswünsche in einem Simulationsszenario
$p(i)$	Wechselwahrscheinlichkeit in die Nachbarzelle i
R_i	Betrag der Bandbreite, die durch die Nachbarstation i reserviert wurde
R	Gesamthöhe der Reservierungen der Nachbarstationen, Summe der R_i
S_{agil}	Bandbreitenschätzung des agilen EWMA-Filters von HoPVarB
S_{extern}	Bandbreitenschätzung des Moduls zur Messung der Zellauslastung
S_{agil}	Bandbreitenschätzung des stabilen EWMA-Filters von HoPVarB
V_{agil}	zu S_{agil} passender Verbindungsfaktor des Moduls zur Messung der Zellauslastung
V_{stabil}	zu S_{stabil} passender Verbindungsfaktor des Moduls zur Messung der Zellauslastung
w	Gewicht eines EWMA-Filters

1 Einleitung

1.1 Motivation und Aufgabenstellung

Mobile Kommunikation ist einer der wesentlichen Trends in der Telekommunikation in den letzten Jahren. Diese Entwicklung findet im rasanten Anstieg der Mobilfunkteilnehmerzahlen ihren Niederschlag, insbesondere in der starken Verbreitung der Mobiltelefonie.

Die Zunahme des Verkehrs in Mobilfunknetzen stellt die Netzbetreiber vor die Herausforderung, die Netzkapazität beständig zu erweitern. Da eine Steigerung der Kapazität einer einzelnen Zelle aufgrund des limitierten Frequenzspektrums nicht möglich ist, kann die Netzkapazität pro Fläche gesteigert werden, indem man größere Funkzellen in mehrere kleinere Zellen mit jeweils der gleichen Kapazität aufteilt. Der Nachteil dieser Erweiterung manifestiert sich in häufigeren Zellenwechseln der Mobilteilnehmer bei aktiver Verbindung (Handover), dessen störungsfreie Abwicklung somit an Bedeutung zunimmt.

Des Weiteren hat auf der Seite der Endgeräte eine starke Diversifikation stattgefunden. Das Spektrum reicht von Mobiltelefonen, die in erster Linie der Sprachkommunikation dienen, über Smartphones bis hin zu Kleinstcomputern wie Personal Digital Assistants (PDAs) oder Notebooks, die primär der Datenkommunikation und –verarbeitung dienen und teilweise ebenfalls Zugang zu Mobilfunknetzen haben. Eine Konsequenz dieser Entwicklungen ist das Zusammenwachsen von Sprach- und Datenkommunikation. Ein Beispiel dafür ist das Universal Mobile Telecomunications System (UMTS), das verschiedenste Dienste integriert und in seinen späteren Releases 4 und 5 auf paketvermittelte Übertragung setzt.

Damit einher geht die Verfügbarkeit einer Vielfalt verschiedener Dienste auf den Mobilgeräten. Neben der klassischen Telefonie sind hier bereits jetzt aus dem Internet bekannte Dienste wie WWW oder Email zu nennen. Mit steigenden Bandbreiten rücken zukünftig auch Multimediadienste wie beispielsweise Streaming Video in das Blickfeld.

Solche Datenströme haben aufgrund der Kompression der Originaldaten die Eigenschaft, dass ihre Bitrate über die Zeit variiert. Somit ist es schwierig vorauszusagen, wie viel Bandbreite (und somit Ressourcen) sie im Netzwerk belegen werden. Solche Prognosen sind jedoch für die Sicherung einer bestimmten Dienstgüte (Quality of Service, QoS), wie sie insbesondere für Multimediadienste sinnvoll ist, von großer Wichtigkeit. Neben Zusagen bezüglich der verfügbaren Bandbreite ist im Mobilfunknetzen die Verfügbarkeit des Dienstes auch im Falle eines Zellenwechsels ein wichtiger Aspekt.

Die vorliegende Arbeit beschäftigt sich mit einem Verfahren, mit dessen Hilfe das Netzwerk auch in den umliegenden Zellen, in die ein Mobilfunkteilnehmer wechseln kann, die vereinbarte Dienstgüte mit hoher Wahrscheinlichkeit zur Verfügung stellt. Um dies sicherzustellen, müssen die umliegenden Basisstationen einen Teil ihrer zur Verfügung stehenden Bandbreite für Handovers reservieren. Im Mittelpunkt steht die Ermittlung der in

einer Funkzelle gegenwärtig von Datenströmen mit variabler Bitrate benutzten Bandbreite. Daraus kann einerseits der Betrag der in Nachbarzellen zu reservierenden Bandbreite abgeleitet werden, und andererseits können Entscheidungen über die Zulassung weiterer Verbindungen getroffen werden.

Dabei stehen sich zwei Ziele gegenüber: einerseits soll das Datenaufkommen auch bei schwankenden Bitraten der einzelnen Verbindungen nicht zu pessimistisch ermittelt werden, um eine gute Auslastung des Netzes zu erreichen. Für diese Art des Verkehrs kommt eine Beurteilung mittels der Peakrate somit nicht in Frage. Andererseits darf die zu reservierende Bandbreite nicht zu gering bemessen werden, da sonst Verbindungen nach Handovers unter Umständen in der neuen Zelle nicht unterstützt werden können. Da diese Zugangskontrolle in ein QoS-Management mittels Differentiated Services (DS) integriert werden soll, ist es ebenfalls wichtig, dass die gute Skalierbarkeit dieses Ansatzes erhalten bleibt. Daher soll die Bandbreitenermittlung aggregiert und nicht pro einzelnem Datenstrom erfolgen. Das Verfahren soll insgesamt einfach und transparent sowie robust gegenüber Fehlern sein. Wichtig ist auch, dass Datenströme mit stark schwankender Bitrate bei der Entscheidung über die Zulassung zum Netzwerk nicht aufgrund ihrer hohen Peakrate benachteiligt werden.

1.2 Aufbau der Arbeit

In Kapitel 2 werden die für das Verständnis der weiteren Arbeit benötigten Grundlagen gelegt. Neben Mobile IP werden die Konzepte der Handoverpriorisierung, der messungsbasierten Zugangskontrolle und des DiffServ-Ansatz vorgestellt. Kapitel 3 definiert zunächst Anforderungen an das zu entwickelnde Konzept, anhand derer Verfahren zur Handoverpriorisierung und messungsbasierten Zugangskontrolle untersucht werden. In Kapitel 4 wird ein auf die Anforderungen abgestimmtes Verfahren entwickelt und beschrieben, die Bedingungen und Ergebnisse der simulativen Bewertung des neuen Verfahrens werden in Kapitel 5 dargestellt. In Kapitel 6 ergänzt eine abschließende Zusammenfassung die Arbeit.

2 Grundlagen

In diesem Kapitel werden die für das Verständnis der weiteren Arbeit benötigten Grundlagen vermittelt. Die angesprochenen Themengebiete entstammen den Bereichen Mobilität, Sprachkommunikation und Internet, an deren Berührungspunkt diese Arbeit angesiedelt ist.

Da das zu entwickelnde Verfahren im mobilen Umfeld angewandt werden soll, spielt die Verwendung von Mobile IP eine wesentliche Rolle. Daher werden zunächst Ziele, Komponenten und Verfahren dieser Ergänzung zum Internet Protokoll erläutert. Um Handovers und deren Verhältnis zu neu beginnenden Gesprächen näher zu beschreiben, steht die Handoverpriorisierung im Mittelpunkt des darauf folgenden Abschnittes. Ein weiterer Aspekt sind Systeme zur Zugangskontrolle, die Abgrenzung verschiedener Verfahren sowie ihre grundlegenden Komponenten. Mit den Eigenschaften und Anforderungen der Differentiated Services und des Mobile Assured Service im Besonderen beschäftigen sich die beiden abschließenden Abschnitte dieses Kapitels.

In dieser Arbeit werden die Begriffe *Datenstrom* und *Verbindung* synonym verwendet. Somit kann es dazu kommen, dass auch für einen Strom von IP-Paketen der Begriff Verbindung verwendet wird, obwohl es sich bei IP [P81] um ein verbindungsloses Protokoll handelt.

2.1 Mobile IP

Das Internet Protocol Version 4 (IP) [P81] setzt voraus, dass die IP Adresse eines Endsystems seinen Anbindungspunkt an das Netz eindeutig beschreibt. Es muss sich in dem von der Adresse angegebenen Subnetz befinden, damit es gesendete Daten empfangen kann.

Um auch mobile Systeme mittels IP erreichen zu können, werden in [P02] Protokollerweiterungen beschrieben, die ein transparentes Routing von IP Paketen zu mobilen Systemen im Internet ermöglichen.

2.1.1 Aufgaben und Zielsetzung

Wesentliche Aspekte bei der Erweiterung der IP Protokollfamilie um Mobile IP sind dabei:

- Mobilsysteme sollen unter einer gleichbleibenden IP Adresse im Internet erreichbar sein, um mit anderen Endsystemen kommunizieren zu können, obwohl sie die Verbindung zum Netzzugangspunkt auf Ebene der Sicherungsschicht unter Umständen beständig wechseln. Dies bedeutet insbesondere, dass diese Adresse auch dann ihre Gültigkeit behält, wenn das System in ein anderes Subnetz wechselt.
- Um die Integration in das bestehende IP Netzwerk so einfach wie möglich zu gestalten, sollen keine Veränderungen an Endsystemen oder Routern erforderlich sein,

die nicht unmittelbar an der Abwicklung von Mobile IP beteiligt sind und somit keine der im folgenden Abschnitt beschriebenen Funktionseinheiten implementieren. Trotzdem soll eine Kommunikation zwischen mobilen und stationären Endgeräten möglich sein.

- Anzahl und Länge der von Mobile IP versendeten administrativen Nachrichten soll möglichst gering sein. Ein Grund dafür ist, dass Mobilsysteme häufig über eine drahtlose Anbindung an das Internet verfügen, die nur eine relativ geringe Bandbreite und eine erhöhte Fehlerrate aufweist und deshalb möglichst wenig belastet werden soll. Mobilsysteme sind häufig batteriebetrieben, so dass die Minimierung des Energiebedarfes einen weiteren Grund darstellt.
- Alle Pakete, die der Information anderer Knoten über den Aufenthaltsort eines Mobilsystems dienen, müssen vor unbefugten Zugriffen geschützt werden, um die Weiterleitung dieser Informationen an dritte Systeme zu verhindern.

2.1.2 Grundlegende Konzepte

In diesem Abschnitt werden die wichtigsten Begriffe und Konzepte von Mobile IP vorgestellt.

Agenten

Mit Mobile IP wurden drei Agenten als neue Funktionseinheiten eingeführt:

- Der *Mobile Node* ist ein Host oder Router, der seinen Netzzugangspunkt ändert. Der Mobile Node kann seinen Aufenthaltsort wechseln, ohne dass sich seine IP-Adresse ändert. Mittels dieser konstanten Adresse kann er dabei mit anderen Netzteilnehmern kommunizieren, sofern eine Verbindung zum Internet auf Ebene der Sicherungsschicht besteht.
- Der *Home Agent* ist ein Router im heimischen Netzwerk des Mobile Node, der Datenpakete an den Mobile Node weiterleitet und Informationen über den Aufenthaltsort des Mobile Node verwaltet.
- Der *Foreign Agent* befindet sich in dem Netzwerk, in dem der Mobile Host sich gerade aufhält. Während der Mobile Node bei ihm registriert ist, leitet der Foreign Agent für den Mobile Node bestimmte Datenpakete vom Home Agent an das Mobile Node weiter. Für Pakete, die der Mobile Node versendet, kann der Foreign Agent als Standardrouter dienen.

Care-Of-Adresse

Es ist zwischen zwei verschiedenen Adressen zu unterscheiden: Zum einen erhält der Mobile Node eine permanente IP-Adresse in demjenigen Subnetzwerk, in dem er sich normalerweise aufhält. Diese Home-Adresse wird ihm auf die gleiche Weise zugewiesen wie die permanente Adresse jedes anderen nicht beweglichen Netzknotens. Hält er sich nicht im heimischen Netzwerk auf, wird die sogenannte Care-Of-Adresse mit dem Mobile Host assoziiert. Sie beschreibt den momentanen Aufenthaltsort des Mobile Nodes. Wechselt dieser das Subnetzwerk, in dem er sich gerade befindet, so verändert sich seine Care-Of-Adresse. Sie muss dem Home Agent mitgeteilt werden, damit dieser Pakete, die an die Home-Adresse gerichtet waren, an den Mobile Host weiterleiten kann. Die Home-Adresse hingegen bleibt auch bei einem Wechsel des Subnetzes gleich. Mit Ausnahme einiger Funktionen zum Mobilitätsmanagement nutzt er sie als Quelladresse aller Pakete, die er versendet.

Betriebsarten

Um eine leichtere Integration in bestehende Netze zu ermöglichen, unterstützt Mobile IP zwei verschiedene Betriebsarten: während sich bei der ersten der Foreign Agent im Festnetz beispielsweise auf einem Netzzugangsknoten befindet, residiert er in der zweiten Betriebsart direkt auf dem mobilen Endgerät. Der Foreign Agent kann somit als Dienstzugangspunkt aufgefasst werden, es handelt sich nicht um einen physikalischen Netzknoten. Bei der Care-Of-Adresse handelt es sich um die Adresse des Knotens, auf dem sich der Foreign Agent befindet.

In der ersten Betriebsart handelt es sich dabei um einen Router oder Netzzugangsknoten im Festnetz. Dazu muss ein Foreign Agent vom Netzbetreiber auf den betreffenden Knoten installiert werden. Diese Vorgehensweise hat zwei wesentliche Vorteile: einerseits können sich mehrere Mobile Hosts eine Care-Of-Adresse teilen, so wird der knappe Adressraum des Internet Protocol Version 4 entlastet. Andererseits kann die Signalisierung zwischen Foreign Agent und Home Agent über das Festnetz erfolgen. Das hat den Vorteil einer besseren Zuverlässigkeit gegenüber einem drahtlosen Link.

In der zweiten Betriebsart befindet sich der Foreign Agent direkt auf dem mobilen Endgerät. Somit muss die Signalisierung über das drahtlose Interface abgewickelt werden. Außerdem benötigt jeder Mobile Node eine eigene Care-Of-Adresse. Dieser Modus hat jedoch den Vorteil, dass der Netzbetreiber keine Mobile IP Funktionalität zur Verfügung stellen muss. Allerdings wird ein Verfahren zur dynamischen Zuweisung von IP Adressen benötigt, um dem Mobilteilnehmer beim Eintritt in den Netzbereich eine Adresse zuweisen zu können. Hierfür kommt beispielsweise DHCP [D97] in Frage. Auf diese Weise ist der Mobile Node für den Home Agent erreichbar.

Dreiecksrouting

Abbildung 2.1 zeigt das Routing von Paketen an und von einem Mobile Node, der sich nicht in seinem Heimatnetzwerk aufhält. Dabei wird hier von der ersten Betriebsart Gebrauch gemacht.

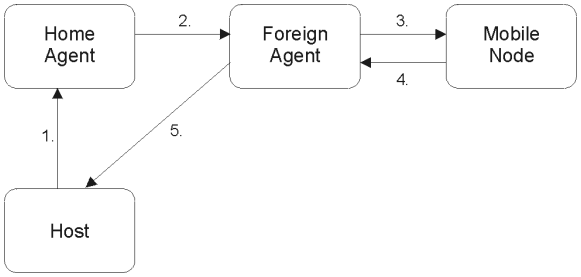

Abbildung 2.1: Dreiecksrouting in Mobile IP

Fünf Schritte sind notwendig:

1. Das Datenpaket an den Mobile Node kommt via normalem IP-Routing im Heimatnetzwerk an, da es die Home-Adresse als Zieladresse trägt. Der Home Agent fängt das Paket ab. Dies ist mit Hilfe des Proxy ARP Mechanismus [P84] möglich.
2. Der Home Agent kennt die Care-Of-Adresse des Mobile Node und muss das Paket dorthin weiterleiten. Dazu wird IP-IP-Tunneling [P96b] verwendet: Das Paket wird in ein weiteres IP-Paket mit der neuen Adresse verpackt und an den Foreign Agent geschickt. Das hat den Vorteil, dass die alte Adresse weiterhin in dem inneren Paket erhalten bleibt
3. Der Foreign Agent entscheidet auf diese Weise, für welchen der bei ihm registrierten Mobile Nodes das Paket bestimmt ist. Er entkapselt das Paket und leitet es an den Mobile Node weiter.
4. Will der Mobile Node ein Paket an einen beliebigen Host verschicken, übergibt er dieses an den Foreign Agent. Als Quelladresse trägt er seine Home-Adresse ein.
5. Der Foreign Agent nimmt die Rolle eines Gateways zum IP-Netzwerk ein. Er dient als Standardrouter für Pakete, die der Mobile Node verschickt. Sie werden mittels Standard-IP-Routing zu ihrem Ziel gesendet.

Die Tatsache, dass Pakete zum und vom Mobile Node durch das Dreiecksrouting unterschiedliche Wege nehmen, führt dazu, dass es auf den beiden Routen zu unterschiedlichen Laufzeiten kommen kann.

2.1.3 Signalisierung

In Mobile IP sind zwei wesentliche Signalisierungen implementiert: das *Agent Discovery* und die *Registrierung*.

Agent Discovery

Mit Hilfe des Agent Discovery stellt der Mobile Node fest, ob er sich in einem Heimatnetzwerk oder in einem fremden Netzwerk befindet. Er kann so auch feststellen, ob er sich von einem Subnetzwerk in ein anderes Netz bewegt hat. Auf diese Weise kann der Mobile Node entscheiden, ob und wo er sich registrieren muss, und im Falle eines Wechsels in ein fremdes Netzwerk die Care-Of-Adresse des neuen Foreign Agent ermitteln.

Es werden zwei Arten von Agent Discovery definiert: Während beim *Agent Advertisement* Home Agent und Foreign Agent ihre Dienste anbieten, fragt bei der *Agent Solicitation* der Mobile Node nach Agenten. Da das Agent Discovery für das Verständnis dieser Arbeit aber nicht von Bedeutung ist, soll an dieser Stelle auf eine detaillierte Beschreibung verzichtet werden, der interessierte Leser sei auf [P02] verwiesen.

Registrierung

Die Mobile IP Registrierung bietet einen flexiblen Mechanismus für Mobile Nodes, um dem Home Agent die aktuelle Lokation mitzuteilen. Darüber hinaus kann er auf diese Weise

- den Dienst der Weiterleitung von Paketen innerhalb eines fremden Netzwerkes beantragen,
- dem Home Agent die momentane Care-Of Adresse mitteilen,
- eine ablaufende Registrierung erneuern und
- sich deregistrieren, wenn der Mobile Node in sein Heimatnetzwerk zurückkehrt.

Wenn sich der Mobile Node außerhalb seines Heimatnetzes befindet, registriert er seine Care-Of-Adresse bei seinem Home Agent. Dies geschieht je nach Betriebart direkt beim Home Agent oder via den Foreign Agent, der die Registrierung weiterleitet. Die Registrierung assoziiert für einen bestimmten Zeitraum die Home Adresse des Mobile Node mit der Care-Of-Adresse des Foreign Agent, danach läuft die Registrierung aus.

Daher ist die Registrierung periodisch zu erneuern. Dies ist typisch für drahtlose, zelluläre Netzwerke, da Zellenwechsel und die hohe Fehleranfälligkeit der Funkstrecke eine zuverlässige Übertragung erschweren. Daher hat die Registrierung nur eine begrenzte Gültigkeit; wird sie innerhalb dieses Zeitraumes nicht erneuert, läuft sie aus. Damit nicht bereits der Verlust einer einzelnen Registrierungsnachricht zu einem solchen Auslaufen führt,

ist die Lebensdauer einer Registrierung dreimal so groß wie die Periodizität des Registrierungsvorganges.

Ablauf des Registrierungsvorganges

Bei der Registrierung werden die Nachrichten *Registration Request* und *Registration Reply* versendet. Abbildung 2.2 zeigt den Ablauf des Registrierungsvorganges in Form eines Weg-Zeit-Diagramms.

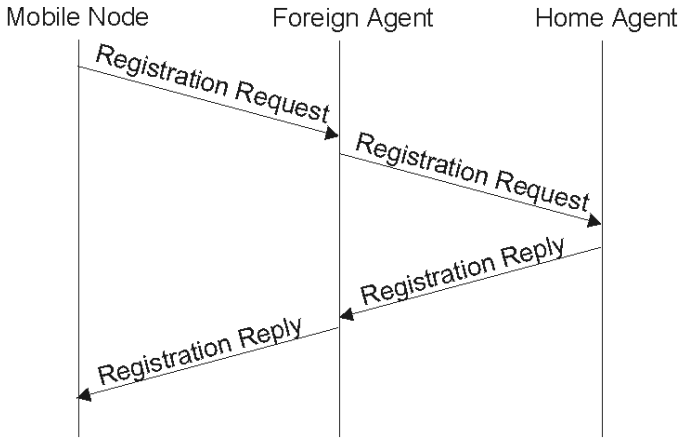

Abbildung 2.2: Ablauf des Mobile IP Registrierungsvorganges

Die Registrierung über den Foreign Agent umfasst 4 Schritte:

Zunächst sendet der Mobile Node einen Registration Request an den Foreign Agent, um den Registrierungsprozess einzuleiten.

Dieser entscheidet über die Registrierung. Lehnt er sie ab (beispielsweise wegen Ressourcenmangels), sendet er direkt einen Registration Reply an den Mobile Node zurück. Ansonsten entnimmt er der Nachricht die Adresse des Home Agent, und leitet sie dann an diesen weiter.

Der Home Agent entscheidet, ob die (Re-)Registrierung korrekt ist und er den Mobile Node unterstützt, und sendet eine Registration Reply an den Foreign Agent zurück. Ein Grund für eine Ablehnung ist zum Beispiel eine fehlgeschlagene Authentifizierung. Im Fall einer positiven Entscheidung wird dann ein IP-IP-Tunnel zwischen Home Agent und Foreign Agent eingerichtet, um an den Mobile Node adressierte Pakete weiterzuleiten.

Der Foreign Agent verarbeitet die Antwort und sendet sie dann an den Mobile Node zurück, um ihn über den Erfolg seiner Registrierung zu informieren.

Die im Rahmen der Registrierung versandten Pakete werden per UDP [P80] verschickt, im Gegensatz zum Agent Discovery wird eine Authentifizierung durchgeführt. Findet die Registrierung direkt beim Home Agent statt, verkürzt sich der Vorgang entsprechend.

2.2 Handoverpriorisierung

2.2.1 Handover in zellulären Funknetzen

Moderne Funknetze sind typischerweise zellulär aufgebaut. Sie bestehen aus einer Anzahl von Basisstationen sowie einem drahtgebundenen Backbone, der die Basisstationen sowohl untereinander als auch mit anderen drahtgebundenen Netzwerken verbindet [SS97]. Das geografische Gebiet, in dem eine Basisstation Mobilfunkdienste anbietet, nennt man *Funkzelle*. Es umfasst häufig nur einen relativ begrenzten Bereich; die Durchmesser reichen von einigen Metern bis zu einigen Kilometern.

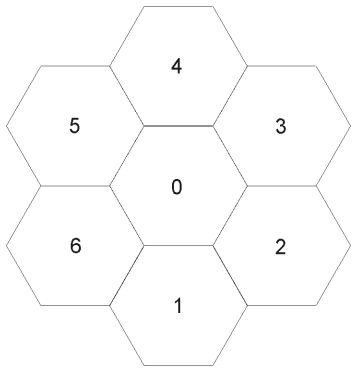

Abbildung 2.3: Hexagonale Anordnung von Funkzellen

Die Anordnung der Funkzellen kann dabei verschieden sein, oft findet man die sogenannte hexagonale Anordnung (Abbildung 2.3). Sie kommt häufig zur Anwendung, da die Näherung der prinzipiell runden Zellen durch Sechsecke zu geringen Überlappungen an den Zellrändern führt. Es kommen aber auch andere Anordnungen wie beispielsweise die eindimensionale Aufreihung von Zellen entlang einer Autobahn in Frage. Deshalb ist es wichtig, dass Verfahren wie beispielsweise Handoverpriorisierungen von der Anordnung der Zellen unabhängig sind.

Verlässt ein Mobilfunkteilnehmer den Bereich der Funkzelle, mit der er momentan verbunden ist, muss eine benachbarte Zelle seine Unterstützung übernehmen. Kommt es während einer aktiven Verbindung zu einem Zellenwechsel, so spricht man von einem *Handover*. Damit ein nahtloser Übergang zwischen den Zellen möglich ist, müssen sie sich an den Rändern überlappen. Dies führt dazu, dass benachbarte Zellen (z.B. Zellen 0 und 2) nicht die gleichen Funkfrequenzen nutzen können, da sie sich sonst im Überlappungsbereich gegenseitig stören würden. Dennoch können Zellen ohne direkte Nachbarschaftsbeziehung die gleichen Frequenzen wiederverwenden (beispielsweise Zellen 2 und 5).

Es lassen sich zwei Zellarten unterscheiden [P96a]: *Makro-* und *Mikrozellen*. Erstere werden in Gebieten eingesetzt, in denen das Funkverkehrsaufkommen pro Fläche gering ist; sie haben einen Durchmesser von mehreren Kilometern. Mikrozellen hingegen messen nur einige

hundert Meter im Durchmesser. Sie werden beispielsweise in Stadtgebieten eingesetzt, in denen das Verkehrsaufkommen pro Fläche so groß ist, dass eine Makrozelle überlastet wäre. Mikrozellen decken nur einen kleinen Bereich ab; Sende- und Empfangsleistung auch der Mobilgeräte können geringer sein, da die Entfernung zur Basisstation kleiner ist.

Wird ein Gebiet sowohl von einer Makro- als auch von mehreren Mikrozellen abgedeckt, spricht man von einem *Overlaynetzwerk*. Der Vorteil liegt darin, dass ein Mobilfunkteilnehmer an einem Ort zwischen zwei unterschiedlichen Basisstationen wählen kann. So können Teilnehmer, die sich nicht bewegen, sich für die Mikrozelle entscheiden, während Teilnehmer mit hoher Mobilität die Makrozelle bevorzugen werden. Durch Overlaynetzwerke ergeben sich komplizierte Modelle für den Handover; sie werden in dieser Arbeit nicht berücksichtigt.

2.2.2 New Call Block und Handover Drop

Ist eine Funkzelle bis an die Kapazitätsgrenze ihrer Basisstation ausgelastet, können keine weiteren Verbindungen in dieser Zelle unterstützt werden: neue Verbindungen müssen von einer Zugangskontrolle abgeblockt werden. Dabei lassen sich zwei Arten von Ablehnungen unterscheiden: *New Call Block (NCB)* und *Handover Drop (HOD)*. Ersteres beschreibt die Zurückweisung eines *New Call*, das heißt, ein Teilnehmer, der in einer Zelle ein neues Gespräch beginnen möchte, wird abgelehnt. Wird im Gegensatz dazu eine Verbindung, die durch einen Zellenwechsel zu der überfüllten Basisstation kommt, abgelehnt, spricht man von einem Handover Drop.

Um die Häufigkeit solcher Verbindungsablehnungen bemessen zu können, werden neben weiteren [P96a] häufig zwei Kennzahlen verwendet. P_{NCB_i} beschreibt die Wahrscheinlichkeit eines New Call Blocks in der Zelle i, wohingegen $P_{HOD_{i,j}}$ die Wahrscheinlichkeit eines Handover Drops bei einem Wechsel von Zelle i in Zelle j benennt. Beide Werte sind wichtig, da diese Wahrscheinlichkeiten bestimmend für die Quality of Service in zellulären Netzwerken sind [SS97].

Prinzipiell gilt, dass die Ablehnungswahrscheinlichkeit für New Calls und Handover Calls gleich ist [SS97], für alle Zellen i und j gilt

$$P_{NCB_i} = P_{HOD_{i,j}} .$$

Dieser Zustand ist jedoch aus Sicht der Mobilfunkteilnehmer unbefriedigend, da der Abbruch einer laufenden Verbindung als störender empfunden wird als eine zeitweilige Zugangsblockade zum Netz. Durch Handover Drops wird die Dienstgüte subjektiv stärker beeinträchtigt. Daher muss es das Ziel des Netzbetreibers sein, die Anzahl der Handover Drops zu beschränken.

Ein Ansatz zur Lösung dieses Problems ist, die Anzahl der Teilnehmer, die gleichzeitig in einer Funkzelle eine aktive Verbindung unterhalten dürfen, zu begrenzen. Dies kann mit Hilfe

einer **Handoverpriorisierung** umgesetzt werden. Das zugrunde liegende Prinzip ist, dass Ressourcen bevorzugt an Handoververbindungen vergeben werden. Auf diese Weise werden sie gegenüber New Calls priorisiert. Dies geschieht, indem man Ressourcen einer Basisstation zweiteilt (Abbildung 2.4).

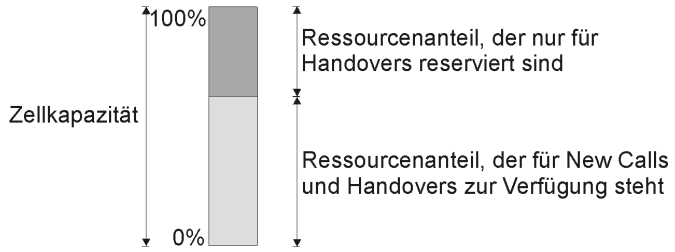

Abbildung 2.4: Reservierung von Ressourcen für Handovers

Einen bestimmten Anteil reserviert man ausschließlich für eingehende Handovers. Der Rest der Ressourcen steht für New Calls, aber auch Handovers zur Verfügung.

Es ist unmittelbar einsehbar, dass durch solche Reservierungen die Zahl der New Call Blocks zunimmt, denn es handelt sich lediglich um eine Umverteilung der Ressourcen zu Gunsten der Handover Calls und nicht um eine Kapazitätserweiterung. Dementsprechend sollte man bei der Festlegung der Menge der reservierten Ressourcen darauf achten, dass P_{NCB} nicht zu groß wird.

2.2.3 Grundlegende Komponenten

Verfahren zur Handoverpriorisierung lassen sich in die Komponenten Bewegungsvorhersage, Ressourcenreservierung und Zugangskontrolle untergliedern (Abbildung 2.5).

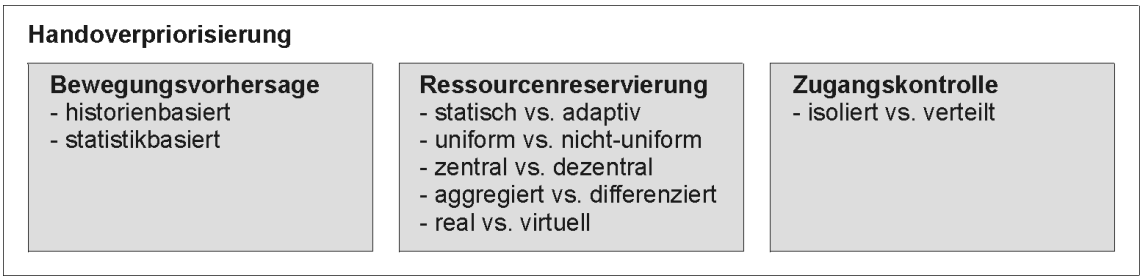

Abbildung 2.5: Komponenten der Handoverpriorisierung

Das Modul zur **Bewegungsvorhersage** hat die Aufgabe, Prognosen über das Bewegungsverhalten der Mobilfunkteilnehmer zu liefern. So sind Aussagen über die Handoverwahrscheinlichkeiten in andere Zellen möglich. Auf Basis dieser Informationen ermittelt die Komponente zur **Ressourcenreservierung** den Betrag der für Handovers zu reservierenden Ressourcen und leitet diese Informationen wenn nötig an die Nachbarzellen weiter. Sie ist somit dafür zuständig, das Verhältnis zwischen für Handovers reservierten

Ressourcen und auch für New Calls zur Verfügung stehenden Ressourcen zu regeln. Aufgabe der **Zugangskontrolle** ist es schließlich, über die Zulassung neuer Verbindungen in der Funkzelle zu entscheiden. Wesentliche Kriterien sind dabei die Zellauslastung und die Größe des für New Calls zur Verfügung stehenden Ressourcenanteils. Grundsätzlich gilt, dass neue Gespräche nur dann zugelassen werden, wenn die aktuelle Auslastung geringer ist als der für New Calls zugängliche Ressourcenanteil.

Nicht alle Verfahren müssen jede dieser Komponenten aufweisen. So könnte man zum Beispiel auf eine Bewegungsvorhersage verzichten und in allen Zellen statisch einen bestimmten Prozentanteil an Ressourcen für Handovers reserviern. Dennoch liegen auch hier Annahmen über das Bewegungsverhalten zugrunde, anhand derer die Größe des Anteils bestimmt wird.

2.2.4 Eigenschaften

Die verschiedenen Komponenten der Handoverpriorisierung können verschiedene Ausprägungen haben, die im Folgenden beschrieben werden. Dadurch lassen sich verschiedene Verfahren dann ähnlich wie in einer Taxonomie ordnen.

Bewegungsvorhersage

Die Bewegungsvorhersage kann auf unterschiedlichen Informationen fußen. Am einfachsten wäre es, den Benutzer vor Verbindungsbeginn über sein Bewegungsverhalten zu befragen. Diese Vorgehensweise hat zwei Nachteile: zum einen sind diese Angaben für den Benutzer lästig; unvorhersehbare Ereignisse wie etwa ein Stau machen sie darüber hinaus unzuverlässig. Eine Möglichkeit, diese Probleme zu umgehen, sind **historienbasierte** Vorhersagen. Sie basieren auf der Annahme, dass das Bewegungsverhalten der Netzteilnehmer in der Vergangenheit Rückschlüsse auf zukünftiges Verhalten ermöglicht. Zellenwechsel der Benutzer werden protokolliert, und die sich daraus ergebenden Bewegungshistorien als Näherung für die Wechselwahrscheinlichkeiten in der Zukunft verwendet. Eine andere Grundlage für Bewegungsvorhersagen sind **statistische Modelle**. Es handelt sich dabei weniger um ein Verfahren, vielmehr sind statistische Annahmen über das Verhalten der Benutzer in die Entwicklung der Handoverpriorisierung eingeflossen. Dieser Ansatz ist einfach, hat aber den Nachteil, unflexibel zu sein: Das Verfahren ist auf bestimmte Annahmen hin optimiert, in unerwarteten Situationen steigt dann die Fehleranfälligkeit stark an.

Ressourcenreservierung

Die Ressourcenreservierungen lassen sich anhand verschiedener Kriterien untergliedern. Zum einen kann man zwischen **statischen und adaptiven Systemen** unterscheiden. Arbeitet eine

Ressourcenreservierung statisch, wird unabhängig vom Verhalten der Nutzer immer ein bestimmter Ressourcenanteil für Handovers reserviert. Sie kommt also ohne eine Bewegungsvorhersage aus. In Situationen mit vielen Handovers wird der Nachteil dieser Vorgehensweise offenbar: der vorgesehene Ressourcenanteil reicht nicht aus, es kommt zu Handover Drops. Alternativ kann eine Ressourcenreservierung adaptiv arbeiten, die Reservierungshöhe hängt vom Bewegungs- und Kommunikationsverhalten der Nutzer ab. In Situationen, in denen viele Handovers erwartet werden, steigt der für sie reservierte Ressourcenanteil an. Vorteile ergeben sich auch, wenn keine Zellenwechsel erwartet werden: Der Ressourcenteil, der im statischen Fall fest reserviert ist, kann nun für New Calls verwendet werden, da die Reservierungen zurückgefahren werden können. Diesen Vorteilen steht jedoch ein erhöhter Aufwand gegenüber: Umliegende Funkzellen müssen über erwartete Handovers informiert werden, damit sie ihre Reservierungen dementsprechend anpassen können.

Weiterhin lassen sich **uniforme und nicht-uniforme Verfahren** unterscheiden. Funktioniert eine Ressourcenreservierung uniform, so ist der Anteil der für Handovers reservierten Ressourcen in allen Funkzellen gleich. Eine nicht-uniforme Vorgehensweise hingegen ermöglicht unterschiedliche Reservierungsniveaus in den unterschiedlichen Zellen. Somit ist es möglich, an sogenannten Hotspots mit besonders vielen Handovers den Reservierungsanteil hochzusetzen.

Eine Unterscheidung der Reservierungsverfahren ist auch anhand der Frage möglich, ob die Berechnung der Reservierungshöhe **zentral oder dezentral** erfolgt. Im ersten Fall ermittelt eine zentrale Einheit, die den Überblick über alle Funkzellen hat, die Reservierungshöhe für die einzelnen Basisstationen. Das hat beispielsweise den Vorteil, dass Bewegungsvorhersagen und Reservierungen auch über den Sichtbarkeitsbereich einer einzelnen Zelle hinaus möglich sind und den Gesamtzustand des Funknetzes einbeziehen können. Allerdings skaliert ein solcher Ansatz nicht gut mit der Netzgröße und der Anzahl der Netzteilnehmer. In großen, stark frequentierten Netzen steigt der Aufwand erheblich an, die zentrale Instanz wird zum Flaschenhals. Diesen Nachteil vermeiden dezentrale, zellen-orientierte Verfahren. In jeder Funkzelle werden die aus den zur Zeit angemeldeten Verbindungen resultierenden Reservierungen lokal ermittelt und an die unmittelbaren Nachbarzellen weitergeleitet. Die verminderte Konzentration des Berechnungsaufwandes wird jedoch mit einer eingeschränkteren Sicht auf das Bewegungsprofil der Verbindungen erkauft: einer Basisstation ist nur bekannt, von welcher Nachbarstation der Nutzer gekommen ist, und in welche Zelle er sie verlassen hat.

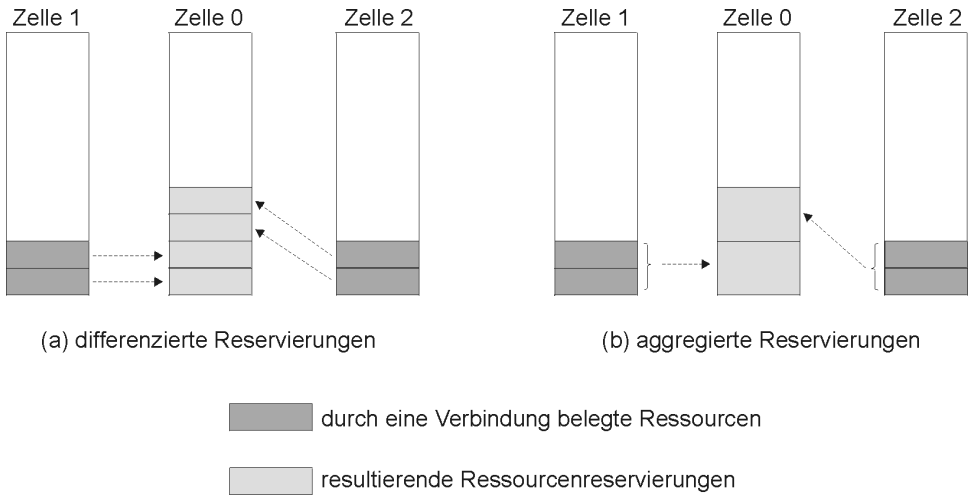

Abbildung 2.6: Vergleich von differenzierten und aggregierten Reservierungen

Innerhalb des zellenorientierten Ansatzes lassen sich **differenzierte und aggregierte Reservierungen** unterscheiden. Im ersten Fall wird für jede einzelne Verbindung eine Bewegungsanalyse durchgeführt, und die daraus resultierenden Reservierungen vorgenommen (Abbildung 2.6 (a)). Das hat zur Folge, dass der Berechnungs- und Kommunikationsaufwand mit der Anzahl der Teilnehmer zunimmt. Im Gegensatz dazu kann man auch die Gesamtheit der Verbindungen einer Zelle betrachten und die Reservierungen aggregiert vornehmen (Abbildung 2.6 (b)). Der erste Ansatz hat den Vorteil, dass zwar prinzipiell genauere Vorhersagen und somit Reservierungen möglich sind, allerdings wächst der Berechnungs- und Datenhaltungsaufwand dabei mit zunehmender Nutzerzahl stark an, die Skalierbarkeit ist schlechter.

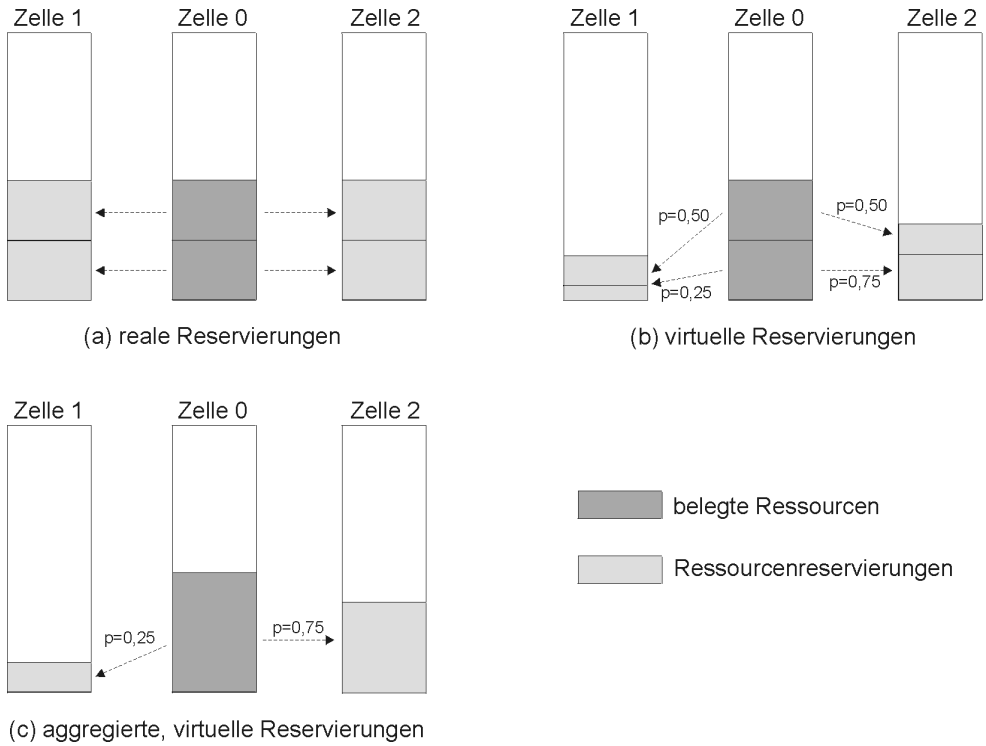

Abbildung 2.7: Vergleich von realen und virtuellen Reservierungen

Abschließend lassen sich **reale und virtuelle Reservierungen** differenzieren. Bei der realen Ressourcenreservierung wird jedem Teilnehmer garantiert, dass die von ihm benötigten Ressourcen im Falle eines Zellenwechsels weiterhin zur Verfügung stehen. Da aber nicht sicher bekannt ist, in welche Nachbarzelle der Nutzer wechselt, muss in allen Nachbarzellen der entsprechende gesamte Betrag reserviert werden (Abbildung 2.7 (a)). Da der Teilnehmer aber nur in eine der Nachbarzellen wechseln kann, werden die Reservierungen in allen anderen Zellen umsonst vorgenommen. Insgesamt werden zu viele Ressourcen reserviert. Bei virtuellen Reservierungen hingegen wird nur der Anteil der belegten Bandbreite in jeder Nachbarzelle reserviert, der der Wechselwahrscheinlichkeit in die betreffende Zelle entspricht (Abbildung 2.7 (b)). Betrachtet man nur eine einzelne Verbindung, führt diese Vorgehensweise dazu, dass im Falle eines Wechsels in der betroffenen Zelle zu wenig Ressourcen reserviert werden, während die Reservierungsanteile in den anderen Zellen nicht mehr benötigt erforderlich sind. Geht man jedoch davon aus, dass in jeder Zelle viele Verbindungen aktiv sind, gleichen sich Über- und Unterreservierungen der jeweiligen Teilnehmer aus. Insgesamt betrachtet wird im Mittel genau das Maß der notwendigen Reservierungen getroffen. Wie in Abbildung 2.7 (c) dargestellt, kann man virtuelle Reservierungen auch aggregiert vornehmen: In jeder Nachbarzelle werden so viele Ressourcen reserviert, wie es dem Produkt aus belegten Ressourcen und der Wechselwahrscheinlichkeit der Gesamtheit aller Teilnehmer in die betreffende Zelle entspricht. Die differenzierte Betrachtung virtueller Reservierungen ist hingegen kaum sinnvoll, da die Virtualität gerade davon profitiert, dass sich verschiedene Verbindungen die Ressourcen teilen.

Zugangskontrolle

Auch für die Zugangskontrolle lässt sich ein Differenzierungsmerkmal anführen: sie kann **lokal oder verteilt** arbeiten. Im ersten Fall hängt die Zulassungsentscheidung nur von der Zelle ab, in die die neue Verbindung den Zugang erbittet. Das hat den Vorteil, dass die Zugangskontrolle keine Informationen über die benachbarten Zellen oder das gesamte Netz benötigt. Eine verteilte Zugangskontrolle hingegen berücksichtigt nicht nur den Zustand der aktuellen Zelle, sondern auch Nachbarzellen oder das gesamte Netz. So könnte man eine neue Verbindung beispielsweise nur dann zulassen, wenn nicht nur in der aktuellen Zelle genügend freie Ressourcen bereitstehen, sondern auch die Nachbarzellen die durch die neue Verbindung ausgelösten Reservierungen vornehmen können.

2.2.5 Handoverpriorisierung für Anwendungen mit variabler Bitrate

Die Reservierung von Ressourcen und die Entscheidung über die Zulassung einer neuen Verbindung hängen maßgeblich von der aktuellen Ressourcenauslastung ab. Eine wesentliche Ressource ist dabei die Bandbreite, die eine Basisstation insgesamt für alle Verbindungen zur Verfügung stellen kann. Daher ist es notwendig, dass eine Basisstation jederzeit weiß, wie viel der verfügbaren Bandbreite im Moment belegt ist. Eine weitere wichtige Information ist, wie viel Bandbreite eine neue Verbindung belegen wird.

Insbesondere wenn in einer Zelle Verbindungen mit variabler Bitrate aktiv sind, ist es schwierig, den Betrag der belegten Bandbreite sicher zu bestimmen. Da die Datenrate der einzelnen Verbindungen schwankt, verändert sich auch die Bandbreite, die sie insgesamt belegen, über die Zeit.

Ein Ansatz zur Ermittlung der aktuell belegten Bandbreitenressourcen ist aus der Zugangskontrolle in drahtgebundenen Netzen bekannt: Bei der messungsbasierten Zugangskontrolle wird die momentan belegte Bandbreite anhand von Messungen ermittelt. Dieser Ansatz soll im folgenden Abschnitt vorgestellt werden.

2.3 Zugangskontrolle

Mechanismen der Zugangskontrolle (Admission Control) haben die Aufgabe, sicherzustellen, dass ein neuer Datenstrom nur dann in ein Netzwerk mit limitiert verfügbaren, gemeinsam genutzten Ressourcen zugelassen wird, wenn dieser die Einhaltung von vorherigen Dienstgütezusagen nicht beeinträchtigt [JSD97]. Solche Zusagen können dabei insbesondere auch Garantien bezüglich der verfügbaren Bandbreite umfassen.

2.3.1 Abgrenzung von parameterbasierten und messungsbasierten Verfahren

Es lassen sich zwei Gruppen von Zugangskontrollverfahren unterscheiden: die parameterbasierten und die messungsbasierten Algorithmen.

Parameterbasierte Verfahren nutzen für die Entscheidung über die Zulassung eine a priori Spezifikation des neuen Datenstromes anhand von Parametern. Dafür kommen zum Beispiel die stochastische Verteilung der benötigten Bandbreite oder die Token-Bucket-Parameter [FH98] in Frage. Aus diesen Angaben, die für alle bereits zugelassenen Ströme bekannt sind, wird die aktuelle Ressourcenauslastung berechnet. Ein neuer Strom wird nur dann in das Netzwerk gelassen, wenn die noch freien Ressourcen die Zulassung möglich erscheinen lassen.

Dieses Vorgehen hat verschiedene Nachteile [GT97a]. Häufig ist es schwierig, den Datenstrom im Vorfeld exakt zu beschreiben. Daher werden häufig Worst-Case-Szenarien zur Spezifikation verwendet, die der Strom zwar auf jeden Fall einhalten kann, die aber zu einer schlechten Netzwerkauslastung führen können, da der Ressourcenbedarf daraufhin überpessimistisch eingeschätzt wird. Da die Netzwerkauslastung durch einfache Kumulation der von den zugelassenen Datenströmen angegebenen Bandbreitenanforderungen ermittelt wird, können eventuell keine weiteren Verbindungen zugelassen werden, obwohl das reale Verhalten der Datenströme dazu führt, dass Ressourcen ungenutzt bleiben.

Messungsbasierte Zugangskontrollverfahren (Measurement-Based Admission Control, MBAC) versuchen, diese Nachteile auszugleichen: Sie berechnen die aktuelle Ressourcenauslastung nicht aus den bei der Zulassung angegebenen Parametern, sondern ermitteln sie aufgrund von periodischen Messungen des Netzwerkzustandes. Ein neuer Strom wird in das Netzwerk gelassen, wenn seine a priori Charakterisierung und die gemessene Ressourcenauslastung in Summe eine Netzüberlastung ausschließen lassen.

Da die Spezifikation des neuen Datenstromes nur noch für seine eigene Zulassungsentscheidung maßgeblich ist, genügen weniger exakte und pessimistische Angaben, um die korrekte Funktion der Verfahren sicherzustellen. Somit wird eine bessere Netzwerkauslastung möglich. Hält ein Datenstrom seine Parameter nicht genau ein, wird dies in den Messungen bemerkt und kann bei der Zulassung weiterer Ströme berücksichtigt werden. MBACs sind parameterbasierten Verfahren somit überlegen, weil sie nicht mit Worst-Case-Verhalten arbeiten müssen, wenn das mittlere Verhalten der Datenströme nicht bekannt ist. Selbst wenn das wirkliche Verhalten bekannt wäre, könnten sich parameterbasierte im Gegensatz zu messungsbasierten Verfahren nicht an Schwankungen im Verhalten der Ströme beispielsweise aufgrund von Selbstähnlichkeit (Long Range Dependence) anpassen [BSJ00].

Allerdings haben die messungsbasierten Verfahren auch Nachteile: Während die parameterbasierten Verfahren auch eine formale theoretische Analyse ihrer Leistungsfähigkeit

anhand der zugrunde liegenden Annahmen und Beschreibung der Datenströme ermöglichen, ist bei den MBACs nur eine simulative Evaluation möglich [BSJ97]. Weiterhin führt die Abkehr von Worst-Case-Bewertungen der von einem Datenstrom belegten Ressourcen zwar zu einer besseren Netzwerkauslastung, doch macht sie definitive Garantien wie beispielsweise eine obere Grenze der Verzögerung nahezu unmöglich [BSJ00]. Somit eignen sich messungsbasierte Zugangskontrollverfahren insbesondere für Dienste, die weniger harte Dienstgütegrenzen als vielmehr Zielvorgaben benötigen.

2.3.2 Komponenten messungsbasierter Zugangskontrolle

Verfahren zur messungsbasierten Zugangskontrolle lassen sich in zwei Komponenten gliedern (siehe Abbildung 2.8).

Abbildung 2.8: Komponenten der messungsbasierten Zugangskontrolle

Das **Messungsmodul** ermittelt den Netzwerkzustand durch periodische Messungen. Zur Ermittlung der aktuell belegten Bandbreite werden die Paketlängen der das Netzwerk passierenden Datenpakete über ein kurzes Intervall S summiert. Teilt man das Ergebnis durch die Intervalllänge S, erhält man die aktuelle Datenrate.

.

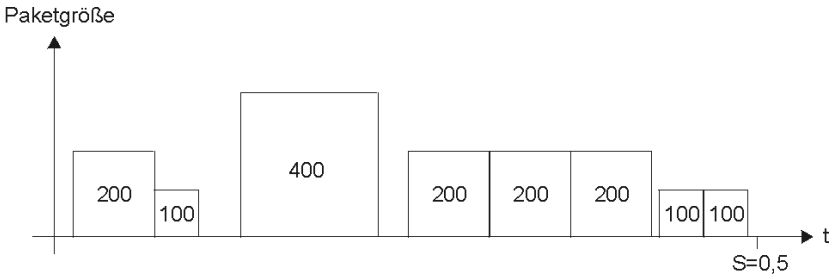

Abbildung 2.9: Berechnungsbeispiel der Datenrate im Intervall S

In dem in Abbildung 2.9 beispielhaft dargestellten Messintervall der Länge S = 0,5 Sekunden werden insgesamt 1500 Byte an Daten übertragen. Es ergibt sich also eine mittlere Datenrate von 3000 Byte/s als Messwert. Nach Ablauf des Intervalls wird die Messung erneut gestartet.

Während bei einigen Verfahren diese punktuellen Messungen das Modell des Netzwerkzustandes widerspiegeln, werden bei anderen Verfahren diese Werte mit Ergebnissen aus der Vergangenheit zu einem Modell des Netzwerkzustandes verknüpft, um

eine Glättung der einzelnen Messwerte zu erreichen. Auf diese Weise kann verhindert werden, dass einzelne Extremausschläge den Gesamtzustand der Messungskomponente verfälschen. Das hat jedoch den Nachteil, dass mit zunehmender Einbeziehung vergangener Messwerte das Modul eine gewisse Trägheit aufweist und so Veränderungen in der Netzauslastung erst nach einer gewissen Zeit bemerkt werden.

Der auf diese Weise ermittelte Netzwerkzustand ist Grundlage für die Arbeit des **Zugangskontrollmoduls**. Mittels für das jeweilige Verfahren charakteristischer Gleichungen wird auf Basis von Netzzustand und der Bandbreitenanfrage entschieden, ob der neue Strom zu einer Ressourcenüberlastung führen würde. In diesem Fall wird dem Datenstrom die Zulassung verwehrt, anderenfalls darf mit dem Senden begonnen werden. Bei einigen Verfahren gibt es auch eine Rückwirkung von vergangenen Zulassungsentscheidungen auf künftige Entscheidungen. Wird beispielsweise eine Verbindung aufgrund einer negativen Zulassungsentscheidung abgewiesen, werden keine neuen Verbindungen angenommen, bevor nicht ein Datenstrom das System verlassen hat.

2.3.3 Taxonomie

Algorithmen zur messungsbasierten Zugangskontrolle lassen sich anhand von drei wesentlichen Kriterien differenzieren (Abbildung 2.10).

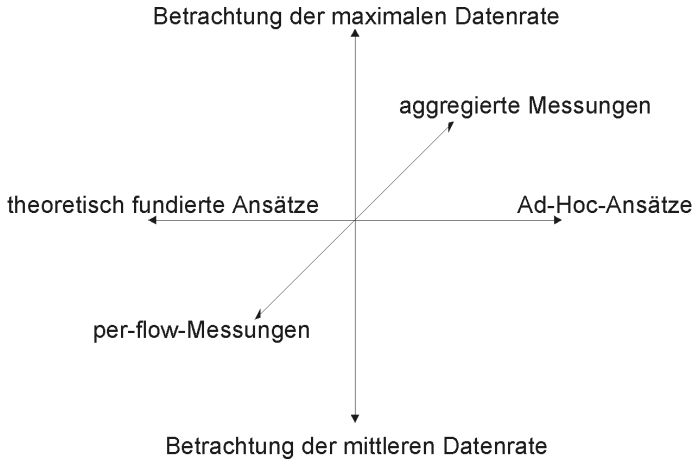

Abbildung 2.10: Differenzierungskriterien für MBACs

Zunächst lassen sich die Verfahren anhand ihres Hintergrundes in **theoretisch fundierte und Ad-hoc-Verfahren** unterteilen [BSJ00]. Erstere basieren auf soliden mathematischen Grundlagen wie beispielsweise der Wahrscheinlichkeitstheorie. Der zum Teil sehr umfassende theoretische Hintergrund führt teilweise zu einer gewissen Intransparenz der Verfahren, weil die Zulassungsentscheidung auf Basis komplizierter Gleichungen gefällt wird. Die Ad-hoc-Verfahren haben keine theoretische Untermauerung, sie funktionieren eher intuitiv.

Ein weiterer Unterschied liegt in der Art und Weise, in der die Messungen durchgeführt werden: **aggregiert oder pro Datenstrom (per-flow)**. Aggregierte Messungen betrachten den Verkehr nicht nach Datenströmen aufgeteilt, sondern kumulieren die Paketgrößen aller Verbindungen. Das hat den Nachteil, dass der Algorithmus nicht weiß, inwieweit die einzelnen Ströme zum Gesamtverkehrsaufkommen beitragen. Diese Informationen können nur mit per-flow Messungen ermittelt werden. Dabei werden die Bitraten der einzelnen Datenströme separat gemessen. Der Aufwand ist deutlich höher, er wächst zudem mit der Anzahl der Verbindungen, so dass die Skalierbarkeit solcher Verfahren eingeschränkt ist.

Als drittes Unterscheidungsmerkmal lässt sich heranziehen, ob die Zulassungsentscheidung anhand der **maximalen Datenrate (Peakrate) oder der mittleren Datenrate (Average Rate)** des neuen Datenstromes gefällt wird. Ein Nachteil des Peakrate-Ansatzes ist, dass es hierbei zu Nachteilen hinsichtlich der Fairness kommen kann, da Ströme mit einer geringeren maximalen Datenrate bevorzugt werden, obwohl sie im Mittel eventuell mehr Ressourcen belegen (siehe auch Abschnitt 3.1.5).

2.4 Differentiated Services

Bei den Differentiated Services (kurz: DiffServ oder DS, [BBC+98], [NBB+98]) handelt es sich um ein Verfahren zur Bereitstellung von Dienstgüte im Internet. Dabei wird der Netzwerkverkehr in verschiedene Dienstklassen unterteilt, die sich in ihren Eigenschaften und Zusicherungen hinsichtlich der Dienstgüte unterscheiden.

Wenn DiffServ genutzt werden soll, vereinbart der Nutzer eine Service Level Specification (SLS) [G02] mit einem Dienstanbieter [XN99]. Sie beschreibt die unterstützten Dienstklassen und die Menge des Verkehrs, der in den einzelnen Klassen erlaubt ist. Man unterscheidet dabei statische und dynamische Spezifikationen. Während statische SLS regelmäßig, also beispielsweise monatlich oder jährlich, vereinbart werden, benötigt die dynamische Spezifikation ein externes Signalisierungsprotokoll, um bei Bedarf den benötigten Dienst anzufordern [XN99].

DiffServ unterscheidet sich von anderen Verfahren zur Sicherstellung von Dienstgüte, insbesondere von den Integrated Services [BCS94], unter anderem durch seine gute Skalierbarkeit. Das bedeutet, dass DiffServ besonders für den Einsatz in großen Netzen mit vielen Datenströmen geeignet ist.

2.4.1 Begriffe und Konzepte

Zunächst sollen einige für DiffServ grundlegende Begriffe geklärt werden.

Datenströme und Klassifizierung

Ein Datenstrom ist eine abstrakte Sammlung von Paketen, die bestimmte Gemeinsamkeiten teilen. Beispiele für Datenströme sind:

- *Mikrostrom (Microflow)*: besteht aus allen Paketen, die von einem gemeinsamen Dienstzugangspunkt der Transportschicht (z.B. einem Socket) kommen und deren gemeinsames Ziel ein anderer Dienstzugangspunkt der Transportschicht ist. Das bedeutet jedoch nicht zwangsläufig, dass alle Pakete des Stromes den gleichen Weg durch das Netz nehmen.
- *aggregierter Strom*: besteht aus einem oder mehreren Mikroströmen, die zusätzliche Gemeinsamkeiten wie beispielsweise die Herkunft von einem gemeinsamen Rechner aufweisen.

Die Zuordnung eines Paketes zu einem bestimmten Strom nennt man *Klassifizierung*. Sie kann anhand eines oder mehrerer Felder im Paketkopf (z.B. Quell- und Zieladresse) erfolgen.

Dienst

Die Definition eines DiffServ-Dienstes besteht aus zwei Komponenten:

1. Eine *Per-hop Behaviour (PHB)* definiert, wie die einzelnen Netzwerkknoten die Pakete des Dienstes weiterleiten sollen. Diese abstrakte Beschreibung wird durch *Mechanismen* wie Buffer Management und Paket Scheduling umgesetzt. Eine solche Vorschrift könnte beispielsweise lauten, Pakete eines bestimmten Stromes immer zuerst weiterzuleiten. Eine Per-hop Behaviour ist somit eine Vorgabe für einen einzelnen, isolierten Netzwerkknoten für die Behandlung einzelner Pakete.
2. Eine *Rule* hingegen stellt sicher, dass die vereinbarten Dienstparameter für den gesamten Datenstrom (ein Mikrostrom oder aggregierter Strom) eingehalten werden können. Eine solche Regel könnte beispielsweise lauten, dass sich niemals mehr als ein Paket eines bestimmten Stromes in der Ausgangswarteschlange eines Routers befinden darf. Der Wirkungsbereich einer Rule erstreckt sich somit auf den gesamten Netzwerkbereich und ist auf Datenströme bezogen.

Für die Definition eines Dienstes sind beide Komponenten unabdingbar. Während die PHB eine geeignete Weiterleitung der Pakete innerhalb des Netzes sicherstellt, sorgen die Rules dafür, dass der Dienst für den gesamten Datenstrom zwischen beliebigen Netzzugangspunkten angeboten werden kann

2.4.2 Architektur

Die Architektur von DiffServ-Netzen basiert auf sogenannten *DiffServ Domains*. Eine solche Domain beschreibt einen zusammenhängenden Netzwerkbereich, der ein gemeinsames Set

von PHBs und Rules umfasst. Allgemein wird das Netz eines Dienstleisters, der Differentiated Services anbietet, aus einer einzigen Domain bestehen.

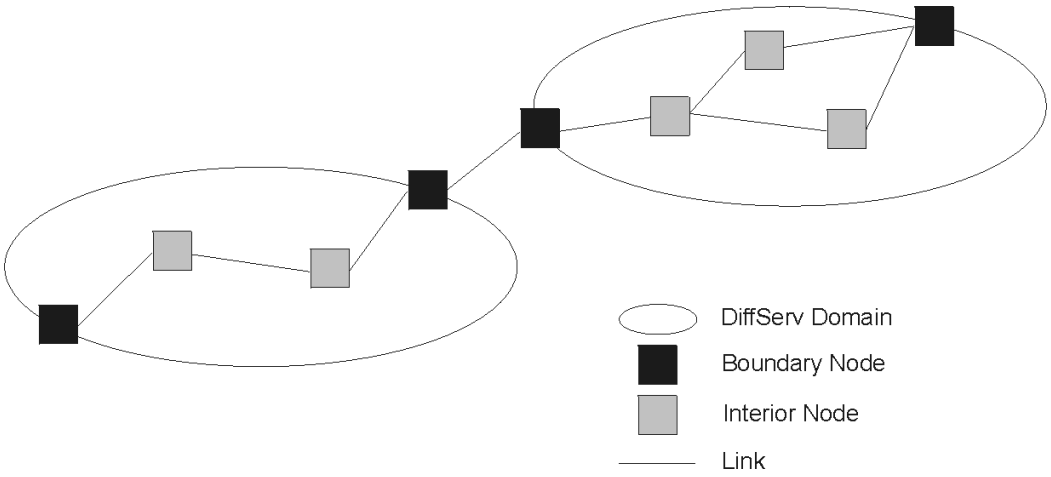

Abbildung 2.11: Beispiel für das Domain-Konzept

Der Zugang zu einer Domain erfolgt über die sogenannten *Boundary Nodes*, Router am Rand des Netzwerkbereiches (Abbildung 2.11). Die Router im Inneren der Domain nennt man *Interior Nodes*. DiffServ weist den beiden Routertypen unterschiedliche Rollen zu: Die Boundary Nodes sorgen für die Einhaltung der Rules, während die Interior Nodes die PHBs umsetzen. Auf die Weiterleitung der Pakete außerhalb der DiffServ Domain kann der Provider keinen Einfluss nehmen. Verbindungen, die mehrere DiffServ Domains durchqueren, sind somit darauf angewiesen, dass sie in allen Domains mit angemessener Dienstgüte unterstützt werden.

Die Weiterleitung von Paketen durch eine DiffServ Domain läuft wie folgt ab:

1. Zuerst gelangt das Paket zu einem Boundary Node. Dieser prüft, ob es mit der SLS harmoniert. Verletzt es diese, wird es verworfen oder mit einer geringeren Priorität transportiert. Dann untersucht der Router unterschiedliche Felder der Paketköpfe der verschiedenen Protokollschichten. Als Ergebnis dieser sogenannten *Multi-field (MF) Classification* setzt der Boundary Node ein bestimmtes Feld im IP Header, den *DS Code Point (DSCP)* [NBB+98]. Er umfasst die sechs Most Significant Bits des TOS Feldes im IP Paketkopf [G02] und legt die PHB fest, mit der das Paket innerhalb der Domain weitergeleitet wird. Alle Pakete mit dem gleichen Code Point bilden ein so genanntes Behaviour Aggregate (BA), dessen Eigenschaften durch Traffic Conditioning und Shaping am Rand der Domain von den Boundary Nodes festgelegt werden, um die Einhaltung der Rules sicherzustellen.
2. Die Interior Node klassifizieren die Pakete, indem sie nur den DS Code Point beachten. Sie führen diese sogenannte BA Classification durch, um zu bestimmen, wie die Pakete weitergeleitet werden. Alle Pakete eines BA werden im Inneren der Domain so behandelt, als gehörten sie zu einem einzigen Strom.

Der Vorteil dieser zweigeteilten Klassifikation ist, dass die komplexe MF Classification nur auf den Boundary Nodes, und nur einmal pro Domain durchgeführt werden muss. Im Inneren der Domain wird von den Interior Nodes nur die viel weniger komplexe BA Classification durchgeführt. Dies erfordert jedoch, dass alle inneren Knoten die gleiche Abbildung von Code Points auf PHB verwenden.

Beispiel: Assured Service

Die Assured Forwarding (AF) PHB [HBW+99] umfasst vier unabhängige Serviceklassen, sogenannte AF-Klassen, mit jeweils bis zu drei unterschiedlichen Dropping Wahrscheinlichkeiten für die Pakete. Innerhalb einer Klasse darf die Reihenfolge von Paketen nicht verändert werden, um Probleme mit den Paketfolgenummern beispielsweise von TCP zu verhindern. Der Assured Service erlaubt den Datenströmen, die vereinbarte Datenrate zu überschreiten. Die Pakete, die außerhalb der SLS liegen, werden mit einem gesonderten Code Point markiert und können von überlasteten Knoten mit einer höheren Wahrscheinlichkeit verworfen werden. Da Bursts erlaubt sind, können sich lange Warteschlangen an Routern bilden, die zu erhöhter Verzögerung und Jitter führen können.

Der Assured Service setzt sich zusammen aus der AF PHB und der Rule, dass Pakete innerhalb der in der SLS vereinbarten Datenrate nur sehr selten aufgrund von überlasteten Knoten verworfen werden sollen. Der Assured Service macht keinerlei Zusagen hinsichtlich der Paketverzögerung. Sie hängen von einem Traffic Conditioning ab, das die Last innerhalb der einzelnen AF-Klassen gering hält. Auch die Einstellung der Parameter der Mechanismen, die die AF PHB implementieren (beispielsweise die Parameter der RED/RIO Warteschlangen Strategie), hat Einfluss auf die Verzögerung.

2.4.3 Eigenschaften

DiffServ hat, insbesondere im Vergleich mit den Integrated Services [BCS94], im Wesentlichen zwei Vorteile: inkrementelle Einführbarkeit und Skalierbarkeit.

- Aufgrund der Trennung von Diensten in PHBs und Rules ist es möglich, als ersten Schritt in Richtung einer Dienstdifferenzierung einige dedizierte PHBs auf jedem Knoten im Netzwerk zusammen mit einfachen Rules auf den Boundary Nodes zu installieren. Dazu reichen eine statische Ressourcenreservierung und Konfiguration zunächst aus. Entwickeln sich die Dienste weiter, können sie die gleichen PHBs unter Umständen in Verbindung mit komplexeren Rules weiterverwenden. Da Rules nur am Rand der Domain relevant sind, müssen nur auf den Boundary Nodes Veränderungen vorgenommen werden; so wird die Umsetzung neuer Dienste einfacher.
- Die gute Skalierbarkeit des DiffServ Ansatzes wird erreicht, indem komplexe Operationen wie die MF Classification und das Traffic Conditioning auf die Boundary

Nodes verlagert wurde. Die Klassifikation im Domaininneren beschränkt sich auf die Unterscheidung von maximal 64 PHBs (6 Bit). Zustandsinformationen müssen nur pro Aggregat, nicht pro Datenstrom gehalten werden. Die Boundary Nodes können komplexere Aufgaben pro einzelnem Datenstrom besser übernehmen, da hier erstens die Anzahl der Ströme insgesamt und zweitens die Anzahl der Pakete pro Sekunde geringer ist.

Dennoch gibt es auch Aspekte, die beim DiffServ-Konzept noch problematisch sind; so sind noch keine Standards für die Dienstanforderung oder dynamische Konfiguration von Knoten vorhanden.

2.5 Mobile Assured Service

Die Handoverpriorisierung und DiffServ stammen aus zwei unterschiedlichen Bereichen der elektronischen Kommunikation. Erstere wurde aufgrund der Bedürfnisse entwickelt, die aus der mobilen Sprachkommunikation entstanden sind und nur bei zellulären Netzen entstehen: Zellwechsel bei laufenden Gesprächen sollten unterbrechungsfrei möglich sein.

Der zweite Bereich, die Datenkommunikation über das Internet, bringt andere Probleme mit sich: der zur Zeit allgemein genutzte Best Effort Service des Internet ist nicht für alle Anwendungen ausreichend, genauere Zusagen über Qualität und Eigenschaften des Dienstes sind beispielsweise für Multimediaanwendungen wie die Übertragung von Ton und Bild über das Internet wünschenswert. Konzepte zur Bereitstellung von Dienstgüte durch DiffServ sind eine Antwort auf diese Frage.

Der Mobile Assured Service stellt den Versuch dar, diese beiden Bereiche miteinander zu verbinden: Es soll eine Dienstgütezusicherung für Datenverkehr über zelluläre Funknetze möglich werden. Der Mobile Assured Service stellt somit eine Weiterentwicklung des in DiffServ definierten Assured Service dar. Eine wesentliche Anforderung an eine solche Zusicherung ist dabei, dass die Verbindung auch nach einem Zellenwechsel die vereinbarten Charakteristika des Dienstes mit hoher Wahrscheinlichkeit einhalten kann. Die Entwicklung einer dafür notwendigen Handoverpriorisierung ist der Gegenstand dieser Arbeit.

2.6 Zusammenfassung

In diesem Kapitel wurden Grundlagen für die weiteren Untersuchungen dargestellt, indem das zugrunde liegende Themengebiet aus verschiedenen Sichtweisen erläutert wurde. Zunächst wurde aufgezeigt, wie mobile Knoten im Internet mittels Mobile IP erreichbar sind. Danach wurde dargelegt, dass die prinzipielle Idee einer Handoverpriorisierung die Verbesserung der Dienstgüte im Mobilfunk durch Reduzierung der Handover Drops zu Lasten der New Call Blocks ist. Die dabei verwendeten Komponenten und ihre Eigenschaften wurden vorgestellt.

Im Hinblick auf Datenströme mit variabler Bitrate wurde ergänzend in Konzepte und Eigenschaften der messungsbasierten Zugangskontrolle eingeführt, was für die weiteren Untersuchungen sehr hilfreich sein wird. Mit dem DiffServ-Ansatz wurde ein gut skalierbares Verfahren zur Bereitstellung von Dienstgüte im Internet erläutert. Der abschließend skizzierte Mobile Assured Service stellt die Klammer um die angesprochenen Themen dar. Er soll die Bereitstellung von Dienstgüte in DiffServ-basierten zellulären Mobilfunknetzen ermöglichen

3 Stand der Forschung

Im vorliegenden Abschnitt wird der aktuelle Stand der Forschung beschrieben, und so eine Grundlage für die darauf aufbauenden Untersuchungen im folgenden Kapitel gelegt. Bereits existierende Verfahren zur Handoverpriorisierung und messungsbasierten Zugangskontrolle werden auf ihre Stärken bzw. Schwächen untersucht. Darüber hinaus wird aufgezeigt, wie man Teile einzelner Verfahren mit anderen kombinieren könnte.

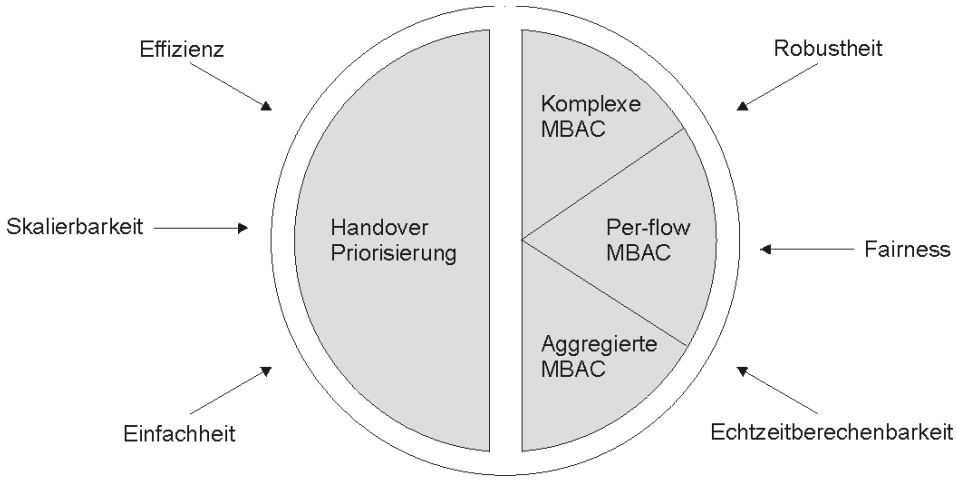

Abbildung 3.1: Betrachtete Bereiche und Anforderungen

Dieses Kapitel ist in verschiedene Abschnitte gegliedert: zunächst werden Anforderungen an einen Algorithmus zur Handoverpriorisierung für Anwendungen mit variablen Bitraten erläutert. Anhand dieser Anforderungen werden verschiedene Verfahren zur Handoverpriorisierung und messungsbasierten Zugangskontrolle (MBAC) betrachtet. Abbildung 3.1 zeigt das Betrachtungsfeld im Überblick. Abschließend werden die Ergebnisse dieses Abschnittes zusammengefasst und bewertet.

3.1 Anforderungen

An einen Algorithmus zur Handoverpriorisierung für Anwendungen mit variablen Bitraten zum Einsatz in einem mobilen DiffServ-Netzwerk sind folgende Anforderungen zu stellen:

3.1.1 Effizienz

Wichtigste Anforderung an das zu entwickelnde Verfahren ist die effiziente Nutzung der verfügbaren Netzwerkressourcen, um einerseits möglichst viele Verbindungen in das Netz aufnehmen zu können, und andererseits einen möglichst hohen Anteil an Handovers unterstützen zu können.

Um eine optimale Netzauslastung zu erreichen, sollten die Reservierungen in den umliegenden Zellen virtuell und adaptiv erfolgen und mit den aktuellen Bewegungsmustern der mobilen Nutzer harmonieren. Weiterhin sollten einzelne Datenströme nur so viele Ressourcen an Übertragungs- und Reservierungskapazität belegen, wie sie tatsächlich benötigen. Insbesondere sollte für Datenströme mit variabler Bitrate nicht ihre Peakrate, sondern eine (evtl. über die Zeit sich verändernde) mittlere Datenrate als Bandbreitenbedarf ermittelt werden.

3.1.2 Skalierbarkeit

Um die gute Skalierbarkeit des DiffServ-Ansatzes, in den die Handoverpriorisierung integriert werden soll, zu erhalten, müssen auch die hier einzusetzenden Verfahren skalierbar sein. Skalierbarkeit bedeutet in diesem Falle, dass Berechnungsaufwand, Kommunikationsaufwand und Größe der notwendigen Datenbasis nicht von der Anzahl der Verbindungen bzw. Nutzer oder der Netzwerkgröße abhängen.

Daraus folgt, dass das Verkehrsaufkommen aggregiert und nicht pro Datenstrom betrachtet werden muss, um den Aufwand von der Zahl der gerade aktiven Verbindungen zu entkoppeln. Darüber hinaus ist es wichtig, dass es eine dezentrale Datenhaltung gibt, deren Größe weitgehend unabhängig von der Anzahl der Nutzer ist. Dies ist erforderlich, um einerseits viele Nutzer unterstützen zu können und andererseits Verkehrskonzentrationen rund um eine zentrale Datenbasis in großen Netzen zu vermeiden. Eine Anordnung der Daten auf den einzelnen Basisstationen hat den Vorteil, dass Anfragen an eine entfernte Datenbank entfallen. Die Kommunikation einer Basisstation muss im Wesentlichen auf eine kleine Anzahl von anderen Stationen (beispielsweise auf die unmittelbaren Nachbarzellen) beschränkt bleiben, damit der Signalisierungsaufwand unabhängig von der Netzwerkgröße bleibt.

3.1.3 Einfachheit

Die Anforderung der Einfachheit steht mit den Forderungen nach Skalierbarkeit und Echtzeitberechenbarkeit im Zusammenhang. Somit soll das Verfahren nur einen begrenzten Kommunikations-, Berechnungs- und Datenhaltungsaufwand nach sich ziehen. Darüber hinaus soll durch eine einfach nachvollziehbare und transparente Funktionsweise die Akzeptanz der Netzwerkoperatoren gegenüber einem solchen Verfahren erhöht werden. Eine klare Steuerbarkeit durch wenige Parameter, deren Auswirkungen auf die Ergebnisse der Handoverpriorisierung klar absehbar sind, ist für den effizienten Betrieb unabdingbar.

3.1.4 Robustheit

Robustheit als Toleranz gegenüber Fehlern und Fehlkonfigurationen ist eine weitere wichtige Eigenschaft, die die Handoverpriorisierung besitzen sollte. Dies ist wichtig, um einerseits Verklemmungen durch Fehlfunktionen einzelner Knoten zu vermeiden und andererseits eine unkomplizierte Handhabung sicherzustellen. Darüber hinaus sollte das Verfahren auch unter verschiedenen Umgebungsbedingungen eine gleichmäßige Leistungsfähigkeit besitzen.

So sollten leichte Veränderungen dieser Bedingungen, wie beispielsweise ein verändertes Bewegungsverhalten der Nutzer oder kurzfristige Lastballungen durch Paketbursts, nicht zu einer wesentlichen Veränderung der Effizienz der Handoverpriorisierung führen. Auch sollte das System den Ausfall einzelner Mobilteilnehmer, zum Beispiel durch eine abreißende Funkverbindung oder eine entleerte Batterie, sowie einzelne oder zeitweilige Paketverluste tolerieren können. Dabei darf es über das isolierte Ereignis hinaus keine Auswirkungen auf das gesamte System oder andere Netznutzer geben. Des Weiteren ist es wünschenswert, dass das Verfahren nicht bereits mit geringfügig vom Optimalwert abweichend eingestellten Parametern stark an Leistungsfähigkeit verliert, da anderenfalls eine Konfiguration sehr schwierig ist.

3.1.5 Fairness

Im Kontext der Zugangskontrolle ist die gleichberechtigte Behandlung aller Zulassungsanfragen wichtig. Es sollte soweit wie möglich vermieden werden, dass Datenströme mit bestimmten Charakteristika bevorzugt oder benachteiligt werden.

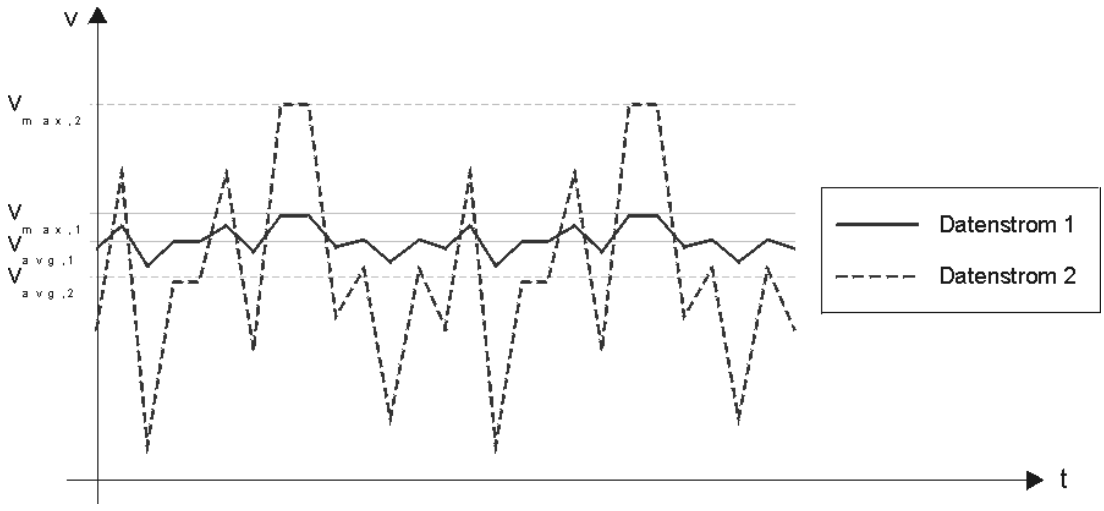

Abbildung 3.2: Maximale und mittlere Datenrate zweier Ströme

Ein Aspekt dieser Problematik sind die Merkmale der Datenströme, die bei der Zulassungsentscheidung berücksichtigt werden. Erfolgt die initiale Bewertung des

Ressourcenbedarfes beispielsweise anhand der Peakrate, kann es sein, dass ein Datenstrom, der weniger burst-artig ist, einem anderen Strom mit gleicher mittlerer Datenrate aufgrund seines geringeren maximalen Bandbreitenbedarfs vorgezogen wird. Ein solcher Fall wird in Abbildung 3.2 verdeutlicht: Während Datenstrom 1 eine niedrigere maximale Datenrate $v_{max,1}$ hat und somit bei einem Peakrate-Ansatz vorgezogen würde, hat Datenstrom 2 eine geringere mittlere Datenrate $v_{avg,2}$ und belegt somit im Mittel weniger Ressourcen. Auch wenn solche Einflüsse unter Umständen nicht vollständig zu vermeiden sind, ist bei der Konzeption einer Handoverpriorisierung darauf zu achten, dass es nicht zu schwerwiegenden Unausgewogenheiten in der Zulassungspraxis kommt.

3.1.6 Echtzeitberechenbarkeit

An Verfahren, die sich adaptiv den Umgebungsbedingungen anpassen sollen, ist die Forderung zu stellen, dass die dafür notwendigen Berechnungen in Echtzeit ausführbar sind. Dies ist erforderlich, damit das System zeitnah auf Veränderungen reagieren kann und der zeitliche Abstand zwischen Ereignissen und daraus resultierender Systemreaktion bei beständig dynamischer Entwicklung nicht beständig zunimmt. Wichtig ist darüber hinaus, dass der Aufwand in Hochlastsituationen nicht überproportional anwächst, um der Anforderung der Skalierbarkeit gerecht zu werden.

Die Grenzen des noch akzeptablen Aufwandes hängen maßgeblich von der Systemumgebung auf denjenigen Netzwerkkomponenten ab, auf denen der Algorithmus zur Handoverpriorisierung laufen muss. Es ist zwar damit zu rechnen, dass die Leistungsfähigkeit der Basisstationen diesbezüglich in der Zukunft zunehmen wird (so werden zum Beispiel die UMTS-Stationen über einen lokalen Speicher verfügen), dennoch wird eine endgültige Beurteilung der Echtzeitberechenbarkeit letztendlich erst im praktischen Einsatz möglich sein. Im Rahmen der theoretischen bzw. simulativen Betrachtung sollte dennoch die Anzahl der erforderlichen Rechenschritte insbesondere in Hochlastsituationen nicht außer Acht gelassen werden.

3.2 Verfahren zur Handoverpriorisierung

In [R01] werden neben einem neuen Algorithmus zur Handoverpriorisierung für Echtzeitanwendungen auch zahlreiche ältere Algorithmen mit ihren Vor- und Nachteilen ausführlich dargestellt. Daher soll an dieser Stelle auf eine differenzierte Analyse der verschiedenen Verfahren verzichtet werden und lediglich der in [R01] von Roth neu entwickelte Algorithmus untersucht werden.

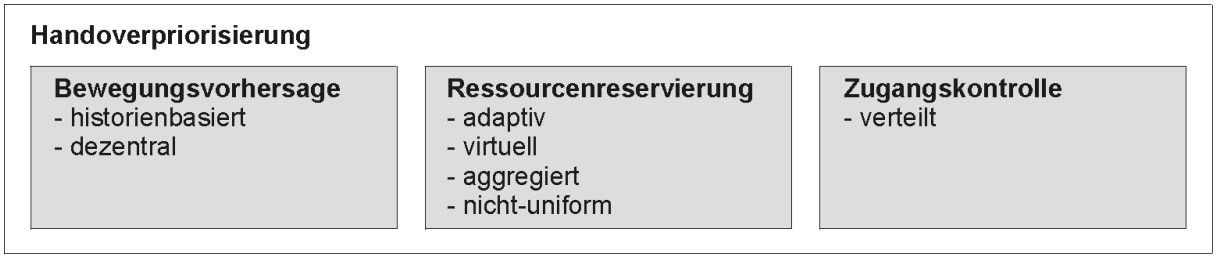

Abbildung 3.3: Komponenten der Handoverpriorisierung in [R01]

Es lassen sich die drei Komponenten Bewegungsvorhersage, Ressourcenreservierung und Zugangskontrolle unterscheiden (siehe Abbildung 3.3). Das Modul zur **Bewegungsvorhersage** arbeitet historienbasiert mit einer dezentralen Datenbasis in jeder Zelle. Sie protokolliert das Bewegungsverhalten der Mobilfunkteilnehmer in der Vergangenheit und leitet daraus Vorhersagen des zukünftigen Verhaltens ab. Wann immer es zu einem Handover kommt, werden die zugehörigen Informationen in Form eines sogenannten Handoverereignisses in die Datenbasis eingetragen. Dabei werden drei verschiedene Verfahren implementiert: Das erste arbeitet nur auf Basis der Zielzelle, es liefert die Wahrscheinlichkeit des Wechsels in eine bestimmte Nachbarzelle nach der Berechnungsvorschrift

$$w(h) = \frac{\left|\{g \mid g \in D \wedge g.Zielzelle = h.Zielzelle\}\right|}{\left|\{g \mid g \in D\}\right|}.$$

Dabei beschreibt *h* das Handover-Ereignis, dessen Wahrscheinlichkeit ermittelt werden soll, und *g* vergangene Handover-Ereignisse in der Datenbasis *D*. Die Wahrscheinlichkeit für einen Handover in die vorgegebene Zielzelle entspricht also dem Verhältnis der Anzahl der Handover-Ereignisse mit derselben Zielzelle zu der Gesamtzahl von Handover-Ereignissen in der Datenbasis.

Die beiden anderen Verfahren sollen an dieser Stelle nicht näher betrachtet werden. Das eine arbeitet zusätzlich auf Basis der Herkunftszelle, allerdings zeigen Simulationsergebnisse in [R01], dass es nur in einem einzigen Szenario Vorteile bietet. Das dritte Verfahren, das auf Basis von Herkunfts- und Zielzelle sowie der Aufenthaltsdauer in der aktuellen Zelle arbeitet, wurde nicht simulativ bewertet.

Im Rahmen der **Ressourcenreservierung** werden periodisch Nachrichten mit den angrenzenden Zellen ausgetauscht. Dabei sendet eine Basisstation den Betrag der in der jeweiligen Nachbarzelle zu reservierenden Ressourcen an die betreffende Zelle. Er wird aus dem Produkt aus aktuell belegter Bandbreite, der Wechselwahrscheinlichkeit in die Zielzelle und dem Steuerparameter Reservierungshöhe (reeller Wichtefaktor zwischen 0 und 1) ermittelt. Bei der Bestimmung der momentanen Bandbreitenauslastung wird ein Peakrate-Ansatz verfolgt, d.h. für jeden Datenstrom wird die im Rahmen der Zugangskontrolle beantragte maximale Datenrate veranschlagt. Die Reservierungen erfolgen adaptiv, das heißt ihre Höhe in den einzelnen Nachbarzellen passt sich dem Bewegungs- und Kommunikationsverhalten der Mobilfunkteilnehmer (und somit der erwarteten Anzahl an

Handovers) an. Aus der Adaptivität ergibt sich unmittelbar der nicht-uniforme Charakter der Reservierungen. Der Reservierungsbetrag wird virtuell berechnet. Das bedeutet, dass nicht in jeder Nachbarzelle die gesamten momentan belegten Ressourcen reserviert werden, sondern überall nur der Anteil, der der Wechselwahrscheinlichkeit dorthin entspricht. Dabei wird aggregiert vorgegangen: Es werden somit nicht für einzelne Verbindungen Reservierungen vorgenommen, sondern die Gesamtheit der belegten Ressourcen betrachtet. Die Nachbarzelle bestätigt den Eingang der Nachricht mit dem Versand einer Antwortnachricht mit Informationen über die dort aktuell verfügbare freie Bandbreite.

Diese Informationen finden Eingang in die **Zugangskontrolle**, die verteilt erfolgt. Bei der Zulassungsentscheidung bezüglich einer neuen Verbindung wird zunächst berücksichtigt, ob die angefragte maximale Bandbreite des Mobilfunkteilnehmers kleiner ist als der Betrag der noch verfügbaren Bandbreite der Basisstation. Diese ergibt sich aus der Gesamtbandbreite abzüglich der Reservierungen und der bereits vergebenen Bandbreite. Ist diese Bedingung nicht erfüllt, wird die Verbindung abgewiesen. Ansonsten wird geprüft, ob die aus der neuen Verbindung resultierenden Reservierungen in den Nachbarzellen aufgrund des Betrags an dort verfügbarer Bandbreite möglich sind. Dabei wird von den Wechselwahrscheinlichkeiten der Datenbasis des Bewegungsvorhersagemoduls ausgegangen. Die neue Verbindung wird nur zugelassen, wenn die aktuelle Zelle die Verbindung unterstützen kann und alle Nachbarzellen ausreichend Bandbreite verfügbar haben, um beim nächsten periodischen Nachrichtenaustausch die entsprechenden Reservierungen vornehmen zu können. In diesem Fall wird die Peakrate der neuen Verbindung zum Betrag der aktuell belegten Bandbreite hinzuaddiert.

Aufgrund der aggregierten Reservierungen, des Nachrichtenaustausches nur mit Nachbarzellen und der dezentralen Datenbasis ist das Verfahren gut skalierbar. Der Autor gibt darüber hinaus an, dass der Algorithmus echtzeitberechenbar ist, da Laufzeit- und Speicherkomplexität unabhängig von der Netzwerkgröße und der Zahl der Verbindungen als konstant angesehen werden können. Das Verfahren besitzt nur einen Steuerparameter und führt keine komplexen Berechnungen durch, somit ist es einfach und transparent. Da der Informationsaustausch und die relevanten Datenaktualisierungen timergesteuert periodisch erfolgen, ist das System äußerst robust gegen Fehler wie zum Beispiel den Ausfall von Mobilteilnehmern.

Allerdings verwendet das Verfahren eine parameterbasierte Zugangskontrolle. Häufig ist es schwierig, a priori den Ressourcenbedarf richtig einzuschätzen. Daher werden diesbezügliche Angaben von der Tendenz her konservativ ausfallen. Da sie nicht nur in die jeweilige Zulassungsentscheidung eingehen, sondern auch folgende Entscheidungen beeinflussen, können sich Fehler durch übervorsichtige Angaben akkumulieren: Geben alle Verbindungen aufgrund konservativer Schätzungen eine zu hohe Datenrate an, können letztendlich weniger Datenströme zugelassen werden. Durch den Peakrate-Ansatz ist bezüglich der Fairness im Hinblick auf Datenströme mit variablen Bitraten mit Nachteilen zu rechnen, weil die

Zulassungsentscheidung solche Verbindungen bevorzugt, die eine niedrige Peakrate bei gleicher mittlerer Datenrate haben. Da das Verfahren die maximale Datenrate für jede Verbindung veranschlagt, wird darüber hinaus insbesondere bei VBR-Strömen, deren mittlere Datenrate weit unter ihrer Peakrate liegt, deutlich mehr Bandbreite in der aktuellen Zelle belegt und in den umliegenden Zellen reserviert, als die Daten wirklich benötigen. Somit muss die Effizienz des Verfahrens hinsichtlich der in dieser Arbeit thematisierten Verkehrsmuster als mangelhaft angesehen werden.

3.3 Verfahren zur messungsbasierten Zugangskontrolle

In diesem Abschnitt werden Verfahren zur messungsbasierten Zugangskontrolle (Measurement Based Admission Control, MBAC) daraufhin untersucht, inwieweit sie sich für den Einsatz in dem zu entwickelnden Handoverpriorisierungsverfahren eignen. Begünstigt durch die Bemühungen, Verfahren zur Bereitstellung von Dienstgüte im Internet zu entwickeln, wurde in den letzten Jahren eine Vielzahl von Arbeiten zum Thema MBAC veröffentlicht. Zunächst werden diejenigen Algorithmen, die bereits wegen der zugrunde liegenden Annahmen ungeeignet sind, in Abschnitt 3.3.1 vorgestellt. Verfahren, die sich wegen des Algorithmus selbst für das vorliegende Szenario nicht oder nur bedingt eignen, werden in den Abschnitten 3.3.2 und 3.3.3 erläutert. Dabei wird aufgezeigt, weshalb ihr Einsatz nicht angebracht erscheint. Prinzipiell geeignete Verfahren werden dann in Abschnitt 3.3.4 genauer besprochen.

3.3.1 Verfahren mit ungeeigneten Annahmen

Zwei Verfahren kommen für den Einsatz im Rahmen einer Handoverpriorisierung bereits wegen der zugrunde liegenden Annahmen nicht in Frage. Das von **Weiss/Dropmann/Godlewski** in [WDG96] vorgeschlagene Verfahren für die Sprachkommunikation basiert auf der Annahme, dass alle Datenquellen im Mobilfunknetzwerk strikte On-Off-Quellen sind und dynamisch nur dann einen Funkkanal zugewiesen bekommen, wenn wirklich gesendet wird. Auf diese Weise werden Funkkanäle nur belegt, wenn gesprochen wird. In Sprechpausen werden keine Daten übertragen, und deshalb auch kein Kanal belegt. Das Verfahren versucht, die Zulassungsentscheidung einer neuen Verbindung davon abhängig zu machen, wie viel Daten zu Beginn der On-Phasen in der Vergangenheit im Mittel verworfen werden mussten, weil nicht rechtzeitig ein Funkkanal zur Verfügung stand. Somit funktioniert das Verfahren nur mit On-Off-Quellen; Datenquellen mit variabler Bitrate können zwar On-Off-Quellen sein, müssen diese Charakteristik aber nicht zwingend aufweisen. Daher kommt dieses Verfahren für die in dieser Arbeit betrachteten Verkehrsmuster nicht in Frage. Man kann zwar auch kontinuierliche VBR-Quellen auf Paketebene als On-Off-Quellen auffassen, indem man das Übertragen eines Paketes als On-Phase betrachtet, und die Pause bis zum nächsten Paket als Off-Phase.

Allerdings scheint die Zuweisung eines Funkkanals für den Versand eines einzelnen Pakets nicht praktikabel: wenn der Funkkanal nicht rechtzeitig zur Verfügung stünde, würde der erste Teil des Paketes fehlen. Während das Fehlen einiger Sprachfetzen zu Beginn einer Redephase bei der Sprachübertragung noch hinnehmbar erscheint, müssen Pakete vollständig übertragen werden. Somit müssten im Falle einer verspäteten Kanalzuweisung die gesamten Daten erneut übertragen werden, was einen großen Overhead erzeugen würde.

Überlagert man verschiedene Datenströme mit variabler Bitrate, so ist die Bitrate des aggregierten Stromes ebenfalls variabel. Dabei lassen sich die Schwankungen in verschiedene Frequenzen untergliedern.

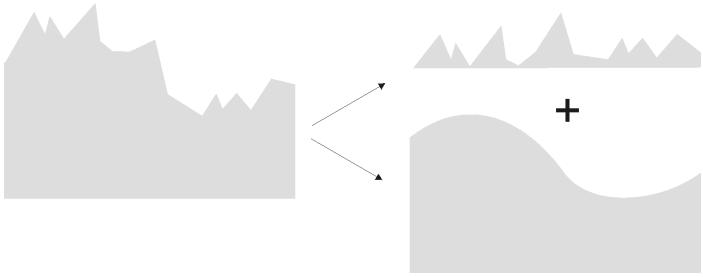

Abbildung 3.4: Unterteilung der Bandbreite in hoch- und tieffrequenten Anteil

Abbildung 3.4 zeigt beispielhaft eine Aufteilung in einen hochfrequenten und einen niederfrequenten Anteil anhand einer Grenzfrequenz. In dem in [GT99] von **Glossauer/Tse** beschriebenen Verfahren wird versucht, anhand des niederfrequenten Anteils eine Zugangskontrolle zu realisieren: Die Schwankungen sollen durch Zulassung bzw. Abweisung oder endende Verbindungen ausgeglichen werden. Zusätzlich wird die Varianz des hochfrequenten Anteils gemessen, um einen Teil der verfügbaren Bandbreite als Puffer für schnelle Schwankungen zu reservieren. Es wird ein Zeitmaßstab $\tilde{T}_h$ ermittelt, so dass:

1. Schwankungen des aggregierten Verkehrs langsamer als $\tilde{T}_h$ durch die (Nicht-) Zulassung von neuen Strömen bzw. den Wegfall von Strömen ausglichen werden und
2. Fluktuationen schneller als $\tilde{T}_h$ per Varianzmessung zur Reservierung von Bandbreite über die mittlere Datenrate hinaus ausgeglichen werden können.

Aus diesen Informationen wird dann die Anzahl der Ströme ermittelt, die maximal zugelassen werden dürfen. Bei den Berechnungen wird vorausgesetzt, dass alle Ströme gleiche Datenraten und Verbindungsdauern haben. In der Realität können die Datenströme aber sehr unterschiedliche Eigenschaften haben, so dass das Verfahren von unrealistischen Bedingungen ausgeht.

3.3.2 Verfahren mit per-flow Messungen

Um der Anforderung der Skalierbarkeit gerecht zu werden, sollte die Messung des Datenaufkommens als Basis für Zulassungsentscheidungen aggregiert erfolgen, da der Aufwand mit zunehmenden Verbindungszahlen bei Messungen pro Datenstrom zu stark

anwachsen würde. Somit sind MBAC-Algorithmen, die auf per-flow Messungen basieren, für die Integration in das zu entwickelnde System prinzipiell ungeeignet. Trotzdem sollen hier einige Verfahren dieser Art in aller Kürze vorgestellt werden.

In [GT97a] beschreiben **Glossauer/Tse** einen Algorithmus, der die Ströme anhand von per-flow-Messungen in Bandbreitenklassen μ_k, k = 1, ..., K, einteilt, wobei μ_k die Klasse der Ströme mit dem Bandbreitenbedarf c_k ist. Dann wird die Bandbreitenverteilung $\{\pi_k\}$ berechnet. Sie stellt die Verteilung des Bandbreitenbedarfes der einzelnen Klassen μ_k dar. Dieser berechnet sich aus der Anzahl der Ströme in Klasse μ_k im Verhältnis zu der Gesamtanzahl der im System befindlichen Ströme, multipliziert mit dem jeweiligen Bandbreitenbedarf c_k. Die Zugangskontrolle basiert auf der Chernoff-Ungleichung, die eine Summe von Zufallsvariablen nach oben hin abschätzt. Mit ihrer Hilfe kann aus $\{\pi_k\}$ die effektive Bandbreite berechnet werden, die eine Grenzbandbreite darstellt, die nur mit einer festlegbaren Wahrscheinlichkeit ε überschritten wird. Ein neuer Strom wird zugelassen, wenn die effektive Bandbreite zuzüglich des neuen Stromes unterhalb der Linkkapazität C bleibt.

Ein weiteres von den beiden Autoren in [GT97b] vorgestelltes Verfahren führt ständig differenzierte Messungen der einzelnen Ströme durch und erlangt so Informationen über deren Bandbreitenerfordernisse. Zum Zeitpunkt einer Zulassungsentscheidung werden aus diesen Informationen der Mittelwert und die Varianz der bereits bestehenden Verbindungen ermittelt. Mit Hilfe des zentralen Grenzwertsatzes wird dann die Anzahl der Verbindungen errechnet, die zugelassen werden dürfen, ohne dass die Wahrscheinlichkeit, dass die verfügbare Kapazität überschritten wird, einen festgelegten Wert übersteigt.

Gibbens/Kelly schlagen in [GK97] insgesamt vier verschiedene MBAC Algorithmen vor, die auf unterschiedlichen Tangenten an die Funktion der effektiven Bandbreite beruhen. Sie stellt eine Grenzbandbreite dar, die von dem aggregierten Datenstrom nur mit einer festlegbaren Wahrscheinlichkeit überschritten wird, und lässt sich mit Hilfe der Chernoff-Ungleichung berechnen. Die Tangenten beschränken dann einen *Annahmeregion (acceptance region)* genannten Bereich. Nur wenn sich die effektive Bandbreite zum Zeitpunkt einer Zugangsanfrage innerhalb der Annahmeregion befindet, wird die neue Verbindung zugelassen. Das Verfahren *Tangent at Arbitary Position* benötigt im Gegensatz zu den anderen Verfahren (siehe Abschnitt 3.3.3) allerdings Messungen der einzelnen Datenströme, da die mittleren Datenraten der einzelnen Verbindungen in die Zulassungsentscheidung einbezogen werden.

3.3.3 Verfahren mit ausgeprägtem theoretischen Hintergrund

In [BSJ00] wurden verschiedene Verfahren zur messungsbasierten Zugangskontrolle auf ihre Leistungsfähigkeit hin untersucht. Es wurde gezeigt, dass Algorithmen mit komplexen Gleichungen zur Messung und Zulassung keine besseren Ergebnisse liefern als einfache

Verfahren ohne theoretische Fundierung. Bereits in [BSJ99] wurde nahegelegt, dass es weniger die Mess- und Zulassungsgleichungen selbst sind, die die Leistung eines Verfahrens ausmachen. Somit qualifizieren sich insbesondere die transparenten und einfach zu berechnenden ad-hoc Verfahren für den Einsatz im Rahmen des in dieser Arbeit zu entwickelnden Algorithmus. Verfahren mit ausgeprägtem theoretischen Hintergrund scheinen wegen ihrer nicht vorhandenen Mehrleistung unvorteilhaft. Dennoch sollen in diesem Abschnitt verschiedene Ansätze dieser Kategorie vorgestellt werden.

Neben *Tangent at Arbitary Position* stellen **Gibbens/Kelly** in [GK97] noch drei weitere Algorithmen vor, die auf Tangenten an die Funktion der effektiven Bandbreite beruhen. In allen drei Verfahren wird die effektive Bandbreite als Funktion der mittleren Bandbreite der Ströme in verschiedenen Verkehrsklassen aufgefasst. Anhand unterschiedlicher Chernoff-Grenzen in den Verfahren ergeben sich verschiedenartige Formen der Funktionen. Somit lassen sich an verschiedenen Stellen Tangenten an die Funktionen anlegen, die die Annahmeregion auf unterschiedliche Weise nach oben beschränken. Bei allen Verfahren gilt: Wird ein Strom abgewiesen, werden keine Ströme in Betracht gezogen, ehe nicht ein Strom das System verlassen hat. In dem Beitrag werden lediglich verschiedene Zulassungskriterien beschrieben; es werden keinerlei Aussagen dazu gemacht, wie die Messungen erfolgen sollen.

Bei dem *Tangent at Origin* Verfahren wird die Tangente im Ursprung der Bandbreitenfunktion angelegt. Das bedeutet, dass die Zulassungsentscheidung sich praktisch nur auf die aggregierten Messungen stützt. So kann es nicht zu Problemen hinsichtlich der Fairness kommen.

Bei dem *Tangent at Peak* Verfahren wird die Tangente am Maximum der Funktion der effektiven Bandbreite angelegt. Auf diese Weise wird das Maximum der maximalen Datenraten aller bereits zugelassenen Ströme in die Zugangsentscheidungen einbezogen. Das Maximum muss bei der Zulassung von den Ströme angegeben werden. Da für alle Ströme nur eine einzige Peakrate angesetzt wird, muss die Zugangskontrolle klassenweise erfolgen, um große Ineffizienzen zu vermeiden. Dabei bilden alle Ströme mit einer ähnlichen Peakrate eine Klasse. Eine Zulassungsentscheidung basiert dann auf einer summarischen Betrachtung aller Klassen. Problematisch ist die Verwendung einer Einflussgröße s, die sich als Minimum einer Summe von rationalen Exponentialfunktionen ergibt. Die Autoren machen keine Aussagen zur Berechnung von s. Verwendet man beispielsweise das Newtonsche Tangentenverfahren, kann sich ein nicht zu unterschätzender Berechnungsaufwand ergeben. Dies gilt insbesondere, weil keine Angaben darüber gemacht werden, bei welchen Veränderungen der Linkauslastung s neu berechnet werden muss.

Das Verfahren *Tangent of Slope One* entspricht dem schon von Floyd in [F96] vorgestellten *Eqivalent Capacity* Verfahren auf Basis der Hoeffding-Grenze, die eine Sonderform der Chernoff-Grenze darstellt. Es wird in Abschnitt 3.3.4 en detail besprochen.

Während alle bisher besprochenen Verfahren auf der mittleren Datenrate beruhen, bezieht das Verfahren von **Reisslein** in [R00] auch höhere Momente ein. Es basiert auf Messungen des

aggregierten Datenstromes. Dazu wird die Zeit in Intervalle aufgeteilt, innerhalb derer die Linkauslastung gemessen wird. Dabei wird eine Historie über die letzten M Messwerte gehalten. Neuere Messwerte werden dabei stärker gewichtet als ältere. Beantragt ein neuer Datenstrom die Zulassung, wird aus diesen Werten die erzeugende Funktion der Bandbreitenverteilung des aggregierten Datenstroms berechnet. Mit Hilfe einer Näherung aus der Theorie der großen Abweichungen wird eine Wahrscheinlichkeit berechnet, mit der es im Fall einer positiven Entscheidung zu Paketverlusten kommt. Liegt diese unter einer festgelegten Grenze ε, wird der neue Strom zugelassen. Dabei wird auch eine Historie über in der Vergangenheit zugelassene Verbindungen geführt und in die Entscheidung einbezogen, so dass diese im Messfenster M noch nicht voll repräsentierten Verbindungen ein größeres Gewicht erhalten. Bei diesem Verfahren wird die Peakrate des zuzulassenden Stroms zur Bewertung herangezogen. Dies ist sowohl im Hinblick auf die Fairness als auch wegen der Überschätzung des mittleren Ressourcenbedarfes der anfragenden Verbindung ungünstig.

Das in [QK01] von **Qiu/Knigthly** vorgestellte Verfahren basiert auf Messungen mit sogenannten M*aximal Rate Envelopes*, die das Verhalten des aggregierten Datenstroms anhand der Messungen als Funktion der Intervalllänge beschreiben. Diese werden genutzt, um bei Ankunft eines neuen Stroms zu prüfen, ob bei einer Zulassung erstens die Verzögerungsanforderungen aller Ströme eingehalten werden können und zweitens die Wahrscheinlichkeit von Paketverlusten einen festgelegten Wert nicht überschreitet. Dies geschieht anhand von wahrscheinlichkeitstheoretischen Überlegungen auf Basis von Gauss-Verteilungen. Das Verfahren hat den Nachteil, dass die Zulassungsentscheidung auf Basis der Peakrate der neuen Verbindung getroffen wird. Somit werden Datenströme, die burst-artig sind, bei der Zugangskontrolle benachteiligt.

3.3.4 Einfache Verfahren mit aggregierten Messungen

Wie bereits in Kapitel 2.3.2 angesprochen, lassen sich Verfahren zur messungsbasierten Zugangskontrolle in die Komponenten Messung und Zugangskontrolle zerlegen, die sich durchaus separat betrachten lassen. Im Folgenden sollen sowohl Verfahren, die nur der Schätzung der Netzwerklast auf Basis von Messungen dienen, als auch komplette Zugangskontrollverfahren beschrieben werden.

Der Point Sample Estimator

Bei dem *Point Sample Estimator* von **Kelly**, der beispielsweise in [BSJ97] beschrieben ist, wird über ein Intervall der Länge S hinweg die Datengröße aller gesendeten Pakete kumuliert. Teilt man die ermittelte Anzahl an Bytes durch die Intervalllänge, erhält man die mittlere Datenrate der letzten S Sekunden. Nach Ablauf des Intervalls wird die Messung erneut gestartet.

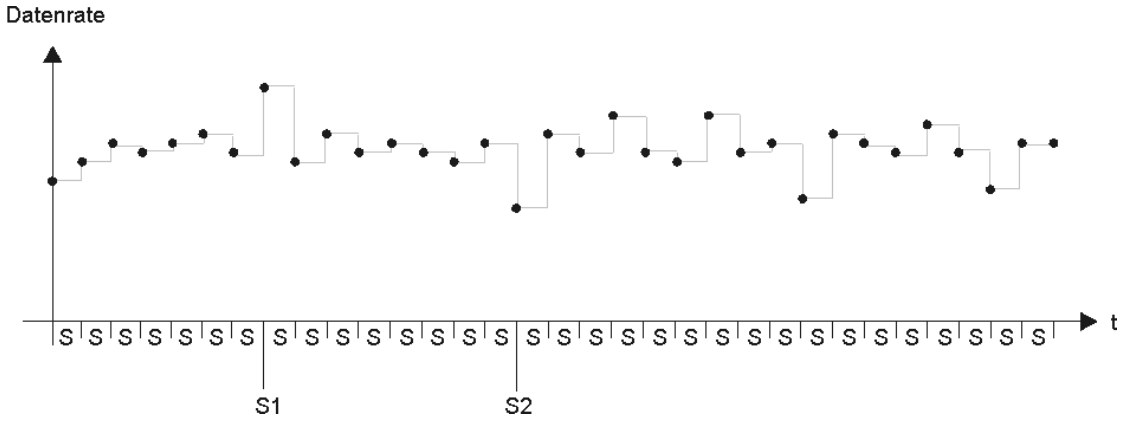

Abbildung 3.5: Messwerte des Point Sample Verfahrens

Im Gegensatz zu anderen Verfahren werden die so aufgenommen Messwerte direkt und ohne weitere Berechnungen als Auslastungsschätzung verwendet. Wie in Abbildung 3.5 zu sehen ist, kann es dadurch zu messergebnisbedingten Schwankungen in der Schätzung kommen. So wird zum Beispiel zum Zeitpunkt *S1* die Netzwerklast überschätzt, während zum Zeitpunkt *S2* eine zu optimistische Einschätzung vorliegt.

Das Verfahren ist einfach; dies führt zu einer positiven Bewertung im Hinblick auf Skalierbarkeit und Berechenbarkeit. Diesen Vorteilen steht eine starke Ungenauigkeit der Messergebnisse als Nachteil gegenüber. Sie rührt von der schlechten Robustheit gegenüber Schwankungen in den Messwerten her. Das wirkt sich letztendlich auch auf die Effizienz negativ aus: Da die Netzauslastungsschätzung wegen der starken Schwankungen nicht verlässlich ist, muss ein Teil der verfügbaren Bandbreite als Puffer unbelegt bleiben, damit es im Falle überoptimistischer Zulassungsentscheidungen nicht zu einer Überbelegung kommt. Das wirkt sich negativ auf die Netzauslastung und somit auf die Effizienz aus.

Das Measured Sum Verfahren

In [JDS+96] beschreiben **Jamin et al.** das so genannte *Measured Sum* Verfahren. Die Zulassungskomponente akzeptiert einen neuen Strom mit der Datenrate D, falls für die aktuelle Netzauslastung $\hat{v}$ und die Linkkapazität *C* gilt:

$$\hat{v} + D < vC \,.$$

Der Wichtefaktor *v* vor der Linkkapazität *C* stellt ein Auslastungsziel dar; die Autoren schlagen einen Wert von 0.9 für *v* vor.

Die aktuelle Netzwerkauslastung wird mittels des *Time Window* Verfahrens geschätzt: Über ein Intervall der Länge *S* wird der Netzwerkverkehr gemessen wie beim Point Sample Estimator. Mehrere Messintervalle werden zu einem Fenster zusammengefasst. Am Ende des Messfensters *T* wird der maximale Messwert (Abbildung 3.6, S1) als Schätzung verwendet.

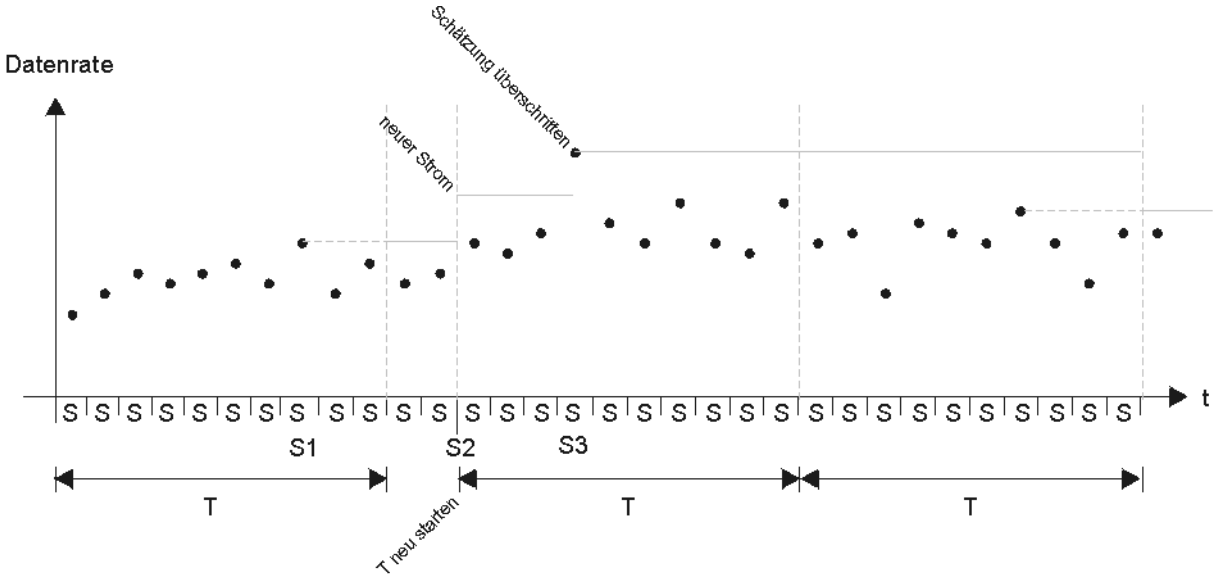

Abbildung 3.6: Das Time Window Messverfahren

Wird ein neuer Datenstrom zugelassen, wird dessen Datenrate zur aktuellen Schätzung hinzu addiert und das Messfenster neu gestartet (S2). Überschreitet in einem einzelnen Intervall der Messwert die aktuelle Schätzung, wird sie auf dessen Wert angehoben (S3).

Ein kleinerer Wert für *S* führt dabei zu höheren Maximalwerten und somit zu einer konservativeren Schätzung, da Bursts sensibler aufgenommen werden. Ist *S* größer, können niedrige Datenraten vor oder nach dem Burst diese ausgleichen, weil sie jetzt noch in dasselbe Messintervall fallen. Die Autoren schlagen für *S* einen Wert von mindestens 100 Paketübertragungszeiten vor.

Ein großer Wert für *T* resultiert in einer längeren Messhistorie. Das führt zu einer konservativeren Schätzung, da es länger dauert, bis ein besonders hoher Messwert (beispielsweise S3) den Schätzwert nicht mehr bestimmt. Für T = S erhält man einen Point Sample Estimator. Um eine statistisch bedeutsame Anzahl von Messwerten zu erhalten, empfehlen die Autoren, $T / S \geq 10$ zu halten.

Die Bewertung des Verfahrens fällt zweigeteilt aus: wie in Abbildung 3.6 erkennbar, ist die Schätzung der Netzauslastung mit dem *Time Window* Verfahren recht konservativ. Dies lässt sich zwar über die Parameter eingeschränkt beeinflussen, ist letztendlich aber prinzipbedingt. Das liegt an der schlechten Robustheit des Messverfahrens gegenüber einzelnen Spitzen in der Datenrate. Somit ist mit einer relativ schlechten Netzwerkauslastung zu rechnen, so dass das Verfahren unter Effizienzgesichtspunkten negativ zu bewerten ist. Andererseits hat das Verfahren den Vorteil, einfach zu sein, was sich positiv auf Transparenz, Skalierbarkeit und Berechenbarkeit auswirkt. Für alle in die Auslastungsschätzung einfließenden Parameter werden klare Vorgaben gemacht.

Die Zulassungskomponente nach dem *Measured Sum* funktioniert intuitiv, ohne komplizierte Berechnungen oder unfangreichen theoretischen Überbau. Dies führt zu einer positiven Bewertung hinsichtlich Einfachheit, Skalierbarkeit und Berechenbarkeit. Zieht man für die Zulassungsentscheidung geschätzte mittlere Datenraten der anfragenden Ströme heran, sind keinerlei Nachteile hinsichtlich der Fairness zu erwarten. Das Verfahren speichert keinerlei Informationen über die einzelnen Ströme. Endet eine Verbindung aufgrund des Ausfalles eines Netzteilnehmers nicht ordnungsgemäß, bleiben somit auch dann keine veralteten Daten zurück. Ebenfalls positiv auf die Robustheit wirkt der grundsätzliche Vorteil von MBACs, dass sich geringe Abweichungen zwischen angegebener und tatsächlicher Datenrate nur auf die betroffene Zulassungsentscheidung, nicht aber auf folgende Entscheidungen auswirken. Der Fokus auf die mittlere Datenrate führt dazu, dass sich die verfügbare Bandbreite gut ausnutzen lässt, so dass die Zulassungskomponente im Gegensatz zur Messkomponente hinsichtlich der Effizienz positiv zu bewerten ist. Nur ein einziger Parameter fließt in die Zulassungsentscheidung ein: Über das Auslastungsziel v kann man Auslastung gegen Paketverluste abwägen; die Autoren schlagen einen sinnvollen Wert vor.

Insgesamt gibt es zwischen der Bewertung der beiden Komponenten einen Gegensatz: Während die Zulassungskomponente positiv zu sehen ist, weist das Messungsmodul zahlreiche Schwächen auf.

Das Equivalent Capacity Verfahren

Floyd stellt in [F96] das *Equivalent Capacity* Verfahren vor, das ebenfalls Messung und Zugangskontrolle umfasst. Für die Zulassungsentscheidung wird die Hoeffding-Ungleichung benutzt, die eine obere Abschätzung einer Summe von Zufallsvariablen aus der Familie der Chernoff-Grenzen ist. Mit ihrer Hilfe kann die äquivalente Kapazität $\hat{C}_H$ berechnet werden, die eine Grenzbandbreite darstellt, die nur mit einer festlegbaren Wahrscheinlichkeit ε überschritten wird. Eine neue Verbindung mit der Peakrate p_{n+1} wird genau dann zugelassen, wenn

$$\hat{C}_H + p_{n+1} \leq C$$

gilt, wobei C die Linkkapazität darstellt. Aus dem Ergebnis der Messung der aggregierten Datenrate $\hat{v}$ und den maximalen Datenraten der Ströme p_i, die bei der Zugangskontrolle angegeben werden muss, ergibt sich

$$\hat{C}_H = \hat{v} + \sqrt{\frac{\ln(1/\varepsilon)\sum_{i=1}^{n}(p_i)^2}{2}}.$$

Wird ein neuer Datenstrom zugelassen, wird seine maximale Datenrate p_i zur aktuellen Lastschätzung $\hat{v}$ hinzuaddiert. Wird ein neuer Strom dagegen abgewiesen, werden keine weiteren Anfragen in Betracht gezogen, bis ein alter Strom das System verlassen hat.

Die Schätzung der Netzwerklast erfolgt mit dem *Exponential Weighted Moving Average (EWMA)* Verfahren: Zunächst werden die Messwerte wie beim Point Sample Verfahren aufgenommen. Um den Nachteil stark schwankender Messwerte zu vermeiden, werden die einzelnen Werte des Datenverkehrs geglättet. Dies geschieht, indem der neue Messwert *M* mit der bisherigen Schätzung $\hat{v}$ verknüpft wird:

$$\hat{v}' = (1-w)\hat{v} + wM \ .$$

Über das Gewicht *w*, das einen Wert zwischen 0 und 1 annehmen kann, lässt sich beeinflussen, wie stark alte Messwerte sich in der aktuellen Netzwerklastschätzung niederschlagen. Hat *w* einen kleinen Wert nahe Null, fließen die aktuellen Messwerte nur zu einem geringen Anteil in das Ergebnis ein. Das hat einerseits den Vorteil, dass Schwankungen in den Messwerten besonders gut geglättet werden. Andererseits kann das System so nur sehr langsam auf Veränderungen reagieren, da es lange dauert, bis sich grundsätzliche Veränderungen in der Netzwerksituation in den gewichteten Ergebnissen niederschlagen. Wählt man für *w* einen großen Wert nahe Eins, verkehrt sich die Situation in das Gegenteil: Aktuelle Messwerte werden stark gewichtet. So ist das System weniger träge, allerdings werden auch messbedingte Schwankungen weniger gut geglättet. Setzt man w = 1, erhält man einen *Point Sample Estimator*.

Abweichend von anderen Verfahren wird für die Zulassungsentscheidung mittels des *Equivalent Capacity* Verfahrens die maximale Datenrate des Stromes herangezogen, was als relativ pessimistisch anzusehen ist und somit unter Umständen Nachteile bezüglich der Effizienz erwarten lässt. Auch bezüglich der Fairness zieht diese Vorgehensweise Nachteile nach sich. Die Berechnung von $\hat{C}_H$ ist relativ komplex: Die Summe der Quadrate der p_i muss zwar nur bei beginnenden oder endenden Verbindungen neu bestimmt werden; der Rest der Wurzel kann ohnehin vorausberechnet werden. Allerdings steigt der Berechnungsaufwand mit der Anzahl der aktiven Verbindungen *n*, wenn auch nur linear. Dennoch ist diese Abhängigkeit nicht zu unterschätzen: würde man die Komponente zur Zugangskontrolle beispielsweise auf einem Base Station Controller anordnen, so kann *n* durchaus in der Größenordnung von einigen 10000 Verbindungen liegen. Somit scheint eine Anordnung auf der einzelnen Basisstation zwingend, trotzdem bleiben Bedenken hinsichtlich der Skalierbarkeit bestehen. Trotz der prinzipiell guten Robustheit gegenüber leichten Abweichungen von der bei der Zugangskontrolle angegebenen Datenrate ergibt sich auch diesbezüglich kein durchgängig positives Bild: Meldet sich eine Verbindung nicht ordnungsgemäß ab, bleibt ihre Peakrate im Verfahren gespeichert, zukünftige Berechnungen der äquivalenten Kapazität werden verfälscht. Die Zugangskontrolle verwendet nur den Parameter zur Verlustwahrscheinlichkeit ε, seine Funktionsweise ist unmittelbar klar. Die Simulationen in [BSJ00] zeigen jedoch, dass ε zu einer sinnvollen Abschätzung der zu erwartenden Verluste nicht geeignet ist.

Positiv ist zu bemerken, dass das *EWMA* Verfahren leicht zu berechnen ist. Es funktioniert intuitiv und transparent. Der Aufwand ist unabhängig von der Anzahl der Verbindungen. Die

einzelnen Messwerte werden geglättet, ohne dass das Verfahren zu pessimistisch ist. Allerdings hängen die Robustheit gegenüber Schwankungen einerseits sowie die Effizienz des Verfahrens andererseits stark von der Wahl des Gewichtes ab. Somit ist die Wahl des Wichtefaktors schwierig, sie stellt immer einen Kompromiss dar.

Das Flip-Flop Verfahren

Eine Lösung für das letztgenannte Problem könnte das von **Kim/Noble** in [KN01] neben anderen EWMA-Filtern beschriebene *Flip-Flop* Verfahren bieten, das ebenfalls auf aggregierten Messungen beruht. Es verwendet zwei EWMA-Filter: einen agilen mit einem Gewicht von 0.9, der wenig träge ist, aber auch die Messungen weniger glättet, sowie einen stabilen mit einem Gewicht von 0.1, der die Messwerte besser glättet, aber träger ist. Beide Filter werden parallel zueinander eingesetzt ((1) in Abbildung 3.7). Welche der beiden Schätzungen nach außen gereicht wird (2), wird von einem Korridor um die aktuellen punktuellen Messwerte gesteuert (4). Normalerweise wird der agilere Filter eingesetzt. Verlassen die Messwerte den Korridor, wird auf den stabileren Filter umgeschaltet (5).

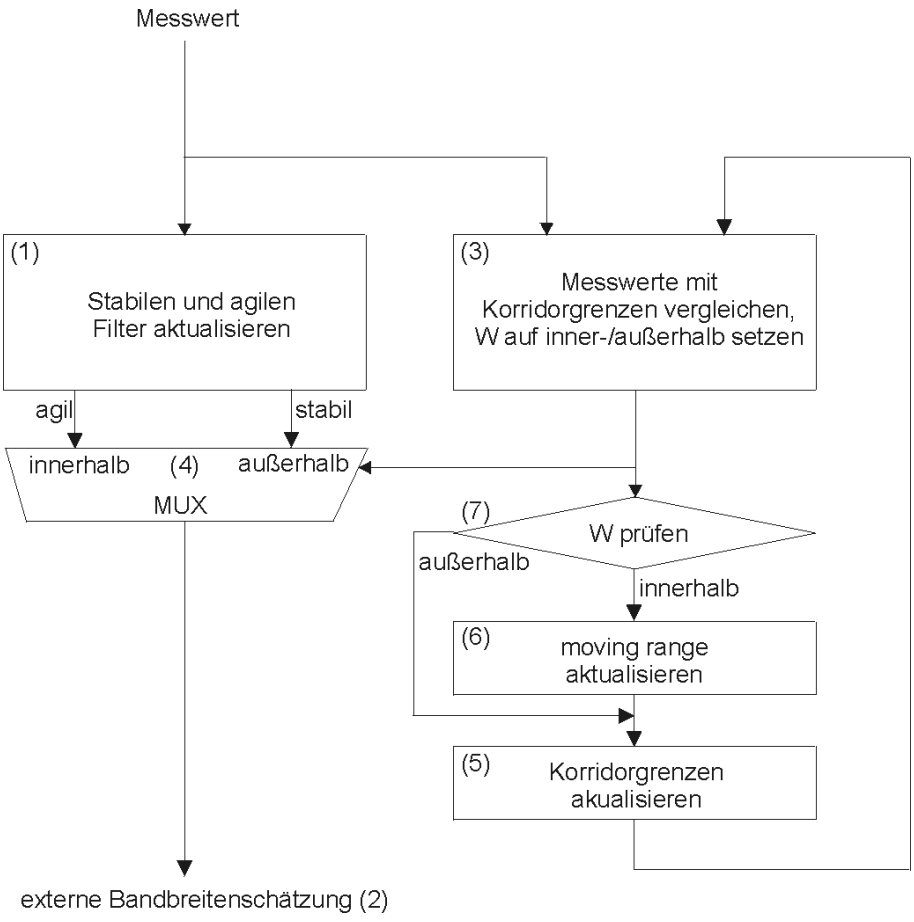

Abbildung 3.7: Schematische Funktionsweise des Flip-Flop-Filters

Der Korridor wird oben und unten durch Grenzwerte beschränkt, die *Upper and Lower Control Limits*. Diese berechnen sich (6) aus dem aktuellen Schätzwert $\hat{v}$ zu

$$\hat{v} \pm 3\frac{\overline{MR}}{1.128},$$

wobei die *Moving Range* $\overline{MR}$ die Differenz zwischen zwei aufeinander folgenden Ergebnissen $|\hat{v}_i - \hat{v}_{i-1}|$ ist, die wiederum durch einen EWMA-Filter mit einem Gewicht von 0.5 geglättet wird (7):

$$\overline{MR} = \frac{\overline{MR}_{alt}}{2} + \frac{|M_i - M_{i-1}|}{2}.$$

$\overline{MR}$ wird allerdings nur dann neu berechnet, wenn der Messwert innerhalb des Korridors lag (8).

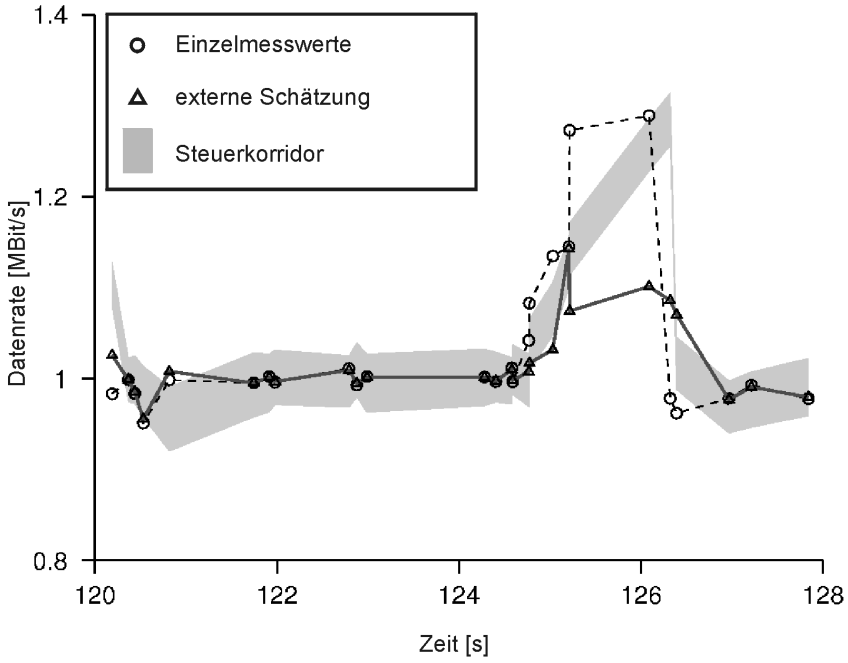

Abbildung 3.8: Messwerte und Schätzungen des Flip-Flop-Filters (aus: [KN01])

Auf Abbildung 3.8 ist zu sehen, wie bei etwa 125 Sekunden die Messwerte den Korridor nach oben verlassen. Daraufhin schaltet das Verfahren auf den stabilen Filter um. Da beide Filter unabhängig von einander arbeiten, fällt die Globalschätzung zunächst, da der stabilere Filter wegen seiner längeren Historie und höheren Trägheit noch niedrigere Werte liefert.

Obwohl deutlich komplexer als ein EWMA-Filter, bleibt der Flip-Flop-Filter mit der intuitiv einsichtigen Korridorsteuerung trotzdem transparent. Der Berechnungsaufwand ist von der Anzahl der Verbindungen unabhängig, so dass von einer guten Skalierbarkeit auszugehen ist. Auch die Echtzeitberechenbarkeit ist somit gegeben. Mit seinen beiden Filtern kann der Algorithmus dabei Veränderungen der Netzwerklast gut in den Schätzungen wiederspiegeln

und so zu einer guten Auslastung beitragen. Bei sprunghaften Messwerten verharrt das Verfahren im stabilen Modus und leistet eine gute Glättung, während bei dicht beieinander liegenden Messwerten der agile Modus eine exakte Wiedergabe der Netzlast bietet. Somit gelingt der Kompromiss zwischen guter Effizienz und guter Robustheit. Die Tatsache, dass zwei Filter verwendet werden, hat darüber hinaus den Vorteil, dass die Wahl der Gewichte weniger kritisch ist als beim EWMA Verfahren, da jedes der Gewichte nur einer Zielsetzung (Robustheit oder Effizienz) gerecht werden muss. Somit können sie besonders ausgeprägt (das heißt 0.9 für den agilen Filter und 0.1 für den stabilen) gewählt werden, wie von den Autoren vorgeschlagen.

3.4 Zusammenfassung

In diesem Kapitel wurden als erstes die Anforderungen an ein Verfahren zur Handoverpriorisierung für Anwendungen mit variablen Bitraten in mobilen DiffServ-Netzen definiert: Effizienz, Skalierbarkeit, Einfachheit, Robustheit, Fairness und Echtzeitberechenbarkeit. Darüber hinaus wurde die Bedeutung der einzelnen Forderungen erläutert.

Anhand dieser Anforderungen wurde dann zunächst ein Handoverpriorisierungsverfahren betrachtet. Dabei wurde auf Stärken hinsichtlich Skalierbarkeit, Einfachheit und Berechenbarkeit hingewiesen. Schwächen bezüglich Effizienz und Fairness entstehen durch den verwendeten Peakrate-Ansatz. Dennoch eignet sich dieses Verfahren als Vorlage für das betrachtete Szenario. Die Robustheit hinsichtlich Fehlern wie dem Ausfall von Netzteilnehmern oder Paketverlusten ist gut, allerdings können sich Fehler bei der Angabe der benötigten Bandbreite auf andere Zulassungsentscheidungen auswirken.

In Abschnitt 3.3 wurden vierzehn messungsbasierte Zugangskontrollverfahren vorgestellt. Zehn Algorithmen sind jedoch aufgrund von ungeeigneten Prämissen, per-flow Messungen oder ihrer Komplexität prinzipiell nicht für den Einsatz im Rahmen des zu entwickelnden Verfahrens geeignet. Darüber hinaus wurde in [BSJ00] gezeigt, dass Algorithmen mit komplexen Gleichungen zur Messung und Zulassung keine besseren Ergebnisse liefern als einfache ad-hoc-Verfahren.

	Messung				Zugangskontrolle	
	Point Sample	Time Window	EWMA	Flip Flop	Measured Sum	Equivalent Capacity
Effizienz	-	-	O	+	+	-
Skalierbarkeit	+	+	+	+	+	-
Einfachheit	+	+	O	O	+	O
Robustheit	-	-	O	+	+	O
Berechenbarkeit	+	+	+	+	+	-
Fairness					+	-

Abbildung 3.9: Vergleich der betrachteten MBACs

Daher wurden vier einfache Mess- und Zulassungsalgorithmen mit aggregierten Messungen in Abschnitt 3.3.4 genauer betrachtet. Abbildung 3.9 stellt die Stärken und Schwächen der Komponenten dieser Verfahren bezüglich der Anforderungen gegenüber. Die Bewertung unterscheidet zwischen gut (+), neutral (O) und schlecht (-). Dabei fällt auf, dass keine der in der Literatur vorgestellten Kombinationen aus Mess- und Zugangskontrollmodul die Anforderungen gut erfüllt. Das Verfahren von Jamin et al. (Time Window/Measured Sum) zeigt Schwächen durch den Einsatz des *Time Window* Messverfahrens, während bei dem von Floyd vorgeschlagenen Algorithmus (EWMA/Equivalent Capacity) die Zugangskontrolle mittels *Equivalent Capacity* Nachteile mit sich bringt. Das *Point Sample* Verfahren (PS) leistet sich wegen der stark schwankenden Messwerte Schwächen hinsichtlich Effizienz und Robustheit, während das *Flip-Flop* Verfahren (Flip Flop) positiv abschneidet. Es beinhaltet jedoch keine Zugangskontrolle.

Trotz der Beobachtungen in [BSJ00] hat die Komplexität der in neueren Arbeiten vorgestellten MBACs eher zugenommen. Bereits in [BSJ99] wurde nahegelegt, dass es weniger die Mess- und Zulassungsgleichungen selbst sind, die die Leistung eines Verfahrens ausmachen. Es wurde vorgeschlagen, den Forschungsschwerpunkt auf die Fairness zu legen. In [BSJ00] wird darüber hinaus angemerkt, dass die Behandlung eingehender und endender Verbindungen einen größeren Einfluss auf die Leistungsfähigkeit eines Algorithmus hat als seine Gleichungen. Dies gilt insbesondere im Umfeld zellulärer Mobilfunknetze, da hier durch Handovers zusätzliche Fluktuationen in der Datenrate entstehen.

Bemerkenswerterweise wird den Aspekten Fairness sowie Einfluss von Verbindungsbeginn/-ende bisher wenig Aufmerksamkeit gewidmet. Nur das Verfahren von Reislein berücksichtigt die Verbindungen gesondert.

Keines der Verfahren ist direkt in der Form, wie es in der Literatur vorgestellt wurde, zur Verwendung im Rahmen der Problemstellung dieser Arbeit geeignet. Damit alle Anforderungen gut erfüllt werden, müssen vielmehr verschiedene einzelne Komponenten neu miteinander kombiniert und um einen Mechanismus zur Berücksichtigung von beginnenden und endenden Verbindungen erweitert werden.

4 Entwurf eines Algorithmus zur Handoverpriorisierung mittels MBAC

In diesem Kapitel wird ein Algorithmus zur Handoverpriorisierung für Anwendungen mit variabler Bitrate in mobilen DiffServ Netzen (HoPVarB) entwickelt und seine Komponenten detailliert beschrieben. Es wird aufgezeigt, dass er den aufgestellten Anforderungen genügt.

4.1 Zielsetzung

HoPVarB lässt sich in die vier Module Bewegungsvorhersage, Ressourcenreservierung, Zellauslastungsmessung und Zugangskontrolle untergliedern (Abbildung 4.1). Um den in Kapitel 3.1 beschriebenen Anforderungen gerecht zu werden, müssen die einzelnen Komponenten bestimmten Voraussetzungen genügen, die im Folgenden beschrieben werden.

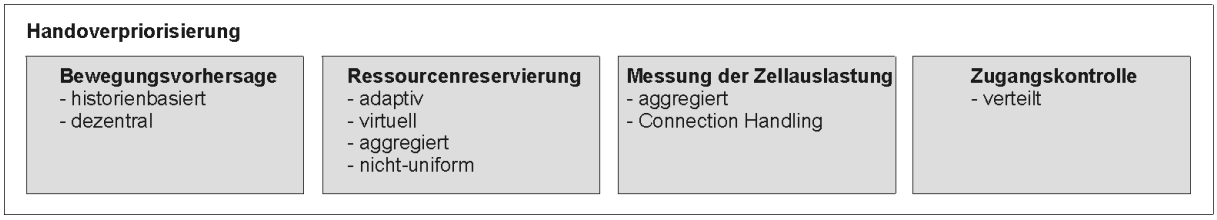

Abbildung 4.1: Komponenten von HoPVarB

Die Bewegungsvorhersage soll historienbasiert arbeiten. Die dafür notwendige Datenbank sollte dezentral angeordnet sein, um die Skalierbarkeit zu gewährleisten. Die Berechnung der Wechselwahrscheinlichkeiten zwischen den Funkzellen soll möglichst einfach sein und so nur einen geringen Berechnungsaufwand verursachen. Durch die Historiendaten wird eine gute Adaptivität der Bandbreitenreservierungen hinsichtlich Anzahl und Verhalten der Netzwerkteilnehmer möglich. Die Reservierungen sollen virtuell, aggregiert und nicht-uniform erfolgen. So wird eine gute Effizienz sichergestellt. Während sich das Verfahren von Roth als Vorlage für die Bewegungsvorhersage anbietet, muss die Komponente zur Ressourcenreservierung im Vergleich zur der in [R01] vorgestellten Vorgehensweise verfeinert werden. Die Reservierungen sollen nicht auf Basis eines Peakrate-Ansatzes erfolgen, sondern sich auf Informationen über die aktuelle Bandbreitennutzung in der Zelle stützen, um die Effizienz zu steigern.

Die Zugangskontrolle soll messungsbasiert und verteilt erfolgen. Es ist erforderlich, dass die periodischen Messungen aggregiert durchgeführt werden, da per-flow Messungen mit zunehmenden Verbindungszahlen nicht skalieren würden. Die einzelnen Messwerte müssen dabei so miteinander verknüpft werden, dass Schwankungen in den Werten geglättet werden, ohne dass die Messungen zu träge werden. In Anlehnung an den Flip-Flop Filter soll hier ein neues Verfahren entwickelt werden: Dabei werden eingehende und endende Verbindungen

gesondert berücksichtigt, um sie schnell in der Netzlastschätzung widerzuspiegeln. Dies ist im mobilen Umfeld besonders wichtig: Während Verbindungen im Festnetz nur durch die Anwendung begonnen oder beendet werden, können in Mobilfunknetzen durch Handovers zusätzlich Verbindungen in einer Zelle beendet und in einer anderen weitergeführt werden. Dadurch kann es zu zusätzlichen Fluktuationen der Ströme kommen. Die Zulassungsentscheidung soll auf Basis der Lastschätzung in der eigenen Zelle und aufgrund von Informationen über freie Ressourcen in den Nachbarzellen erfolgen. Dabei wird die mittlere Datenrate des neuen Stromes als Maßstab herangezogen.

Eine Gewichtung von neu zugelassenen Verbindungen gegenüber unterstützten Handovers soll anhand eines einzigen Steuerparameters möglich sein.

Obwohl verschiedene Parameter Einfluss auf die Funktionsweise von HoPVarB haben, soll sich die Netzauslastung mittels eines einzigen externen Steuerparameters gezielt beeinflussen lassen. Alle anderen Einflussgrößen werden anhand von Umgebungsbedingungen oder aus theoretischen Überlegungen heraus gewählt. Auf diese Weise soll die Administration möglichst einfach gemacht werden.

4.2 Bewegungsvorhersage

Das Modul zur Bewegungsvorhersage hat die Aufgabe, Informationen über die Wechselwahrscheinlichkeiten der Mobilfunkteilnehmer in die benachbarten Zellen zur Verfügung zu stellen. Diese Informationen werden benötigt, um dann in der betreffenden Zelle die entsprechende Ressourcenmenge zu reservieren, um so Handovers in diese Zelle mit großer Wahrscheinlichkeit unterstützen zu können. Der hier vorgestellte Ansatz basiert auf einem der in [R01] verwendeten Vorhersageverfahren. Die zugrunde liegende Annahme ist, dass das Bewegungsverhalten in der näheren Zukunft mit einer großen Wahrscheinlichkeit dem Verhalten in der Vergangenheit ähnelt. Diese Eigenschaft der Bewegungsströme nennt man Selbstähnlichkeit. Kennt man also die Wechselwahrscheinlichkeiten eines vergangenen Zeitraumes, kann man diese als Näherung für zukünftige Wahrscheinlichkeiten verwenden. Diese Vorgehensweise nennt man historienbasiert.

Um Informationen über das Bewegungsverhalten der Mobilfunkteilnehmer zu gewinnen, werden Handovers in andere Zellen protokolliert. Für jeden dieser Zellenwechsel wird ein Datensatz generiert, der Informationen über die Zielzelle enthält. Er wird in einer Datenbasis gespeichert, die auf jeder Basisstation angeordnet ist. Endet ein Gespräch in der aktuellen Zelle, wird auch dieses in der Datenbasis vermerkt.

Auf der Grundlage der Datensätze d in der Datenbasis D kann eine Wahrscheinlichkeitsfunktion p definiert werden. Sie erhält als Argument die Zielzelle z und liefert die Wechselwahrscheinlichkeit nach z zurück:

$$p(z) = \frac{\left|\{d \mid d \in D \wedge d.Zielzelle = z\}\right|}{\left|\{d \mid d \in D\}\right|}.$$

Die Wechselwahrscheinlichkeit entspricht somit dem Verhältnis zwischen der Anzahl der die Zielzelle beschreibenden Datensätze und der Gesamtanzahl der Datensätze in der Datenbasis.

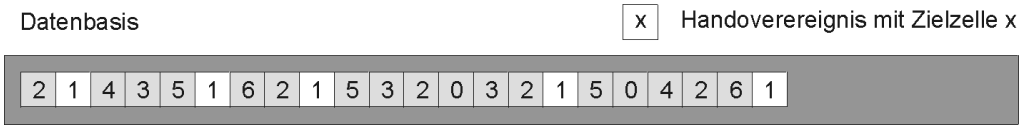

Abbildung 4.2: Beispieldatenbasis, Datensätze mit der Zielzelle 1 sind hervorgehoben

Dieses Verhältnis lässt sich durch bloßes Abzählen ermitteln, wie in Abbildung 4.2 an einem Beispiel deutlich gemacht wird. Betrachtet wird der Datenbestand einer Zelle 0 in der Mitte von 6 Nachbarzellen. Fünf der 22 Datensätze beinhalten die Zielzelle 1, somit gilt für die Wechselwahrscheinlichkeit nach 1:

$$p(1) = \frac{5}{22} = 0{,}2\overline{27} \approx 23\%.$$

Bemerkenswert ist, dass die Summe der Wechselwahrscheinlichkeiten in die 6 Nachbarzellen nicht gleich eins ist. Das liegt daran, dass nicht alle Verbindungen die Zelle aufgrund eines Handovers verlassen, sondern einige Verbindungen in der aktuellen Zelle beendet werden. In einem solchen Fall wird ein Datensatz mit der Zielzellennummer der eigenen Zelle abgespeichert.

Diese Datenbasis hat eine festgelegte Größe, um den Speicherplatzbedarf des Verfahrens an die Gegebenheiten anzupassen. Ist die Kapazitätsgrenze erreicht, verdrängen neu eingefügte Datensätze alte Einträge aus der Datenbasis. Das ist jedoch nicht als Nachteil anzusehen. Es führt vielmehr dazu, dass die Datenbasis immer nur das Bewegungsverhalten eines begrenzten Zeitintervalls repräsentiert. So werden Veränderungen in den Bewegungsmustern erkannt, da die Ereignisse, die noch dem alten Verhalten entsprechen, nach und nach aus der Datenbasis entfallen.

Somit beeinflusst die Größe der Datenbasis auch das Vorhersageverhalten. Mit zunehmender Datenmenge reicht das Gedächtnis der Historie immer weiter zurück. Einerseits geben mehr gespeicherte Ereignisse einen genaueren Eindruck des Bewegungsverhaltens wieder; es ist unmittelbar einsehbar, dass bei beispielsweise 7 Nachbarzellen 300 Ereignisse in der Datenbasis eine genauere Vorhersage ermöglichen als nur drei Ereignisse. Andererseits werden Veränderungen in den Bewegungsmustern erst nach längerer Zeit durch die berechneten Wahrscheinlichkeiten wiedergegeben. Das liegt daran, dass in einer größeren Datenbasis neu hinzukommende Ereignisse im Verhältnis weniger Gewicht haben als in einem kleineren Datenbestand. Es dauert länger, bis die Ereignisse des alten Verhaltens aus der Datenbasis herausfallen.

4.3 Messung der Zellauslastung

Die Komponente zur Messung der Zellauslastung hat die Aufgabe, jederzeit Informationen über den Betrag der momentan belegten Bandbreitenressourcen anzubieten. Diese Informationen werden einerseits von der Ressourcenreservierung benötigt, um die Reservierungshöhe in den Nachbarzellen zu bestimmen, andererseits fließt die Zellauslastung in die Zulassungsentscheidungen der Zugangskontrolle ein.

Die Auslastung der Zelle wird in diesem Verfahren durch Messen der übertragenen Datenmenge ermittelt. Dabei wird in zwei Schritten vorgegangen: Zuerst wird die aktuelle Netzauslauslastung gemessen. Danach wird dieser punktuelle Messwert (Point Sample) mit der bisherigen Schätzung der Netzauslastung verknüpft, um so die einzelnen Messwerte zu glätten. Darüber hinaus werden begonnene oder beendete Verbindungen separat berücksichtigt.

4.3.1 Ermittlung der einzelnen Messwerte

Um die Messwerte zu ermitteln, wird über ein Intervall der Länge S hinweg die Datengröße aller gesendeten Pakete kumuliert. Teilt man die ermittelte Anzahl an Bytes durch die Intervalllänge, erhält man die mittlere Datenrate der letzten S Sekunden.

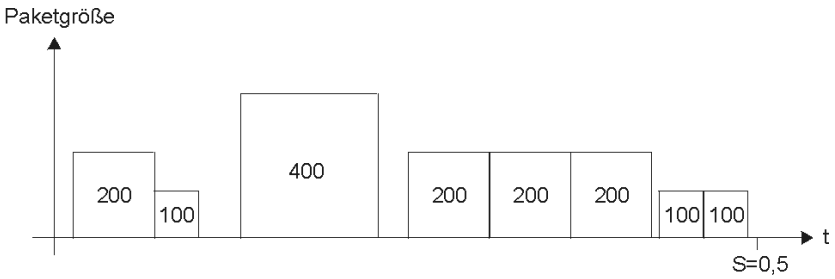

Abbildung 4.3: Berechnungsbeispiel der Datenrate im Intervall S

Zur Verdeutlichung noch einmal das Beispiel aus Kapitel 2.3.2: In dem Messintervall der Länge S = 0,5 Sekunden werden insgesamt 1500 Byte an Daten übertragen (Abbildung 4.3). Es ergibt sich also eine mittlere Datenrate von 3000 Byte/s als Messwert. Nach Ablauf des Intervalls wird die Messung erneut gestartet.

Zur Wahl der Intervalllänge von S: Während ein längeres Intervall genauere Mittelwerte liefert, führt ein kürzeres Intervall zu aktuelleren Werten. In [JDS+96] schlagen die Autoren vor, dass S mindestens 100 Paketübertragungszeiten lang sein sollte.

Laut [GPP02] sind die per UMTS versendeten Pakete maximal 1500 Byte groß. Dies ist gleichzeitig die *Maximum Transfer Unit* (MTU, maximale Paketgröße) der meisten anderen Netzwerktypen. Legt man eine Gesamtbandbreite einer Zelle von ein bis zwei Mbit/s zugrunde, können pro Sekunde 87 bis 174 Pakete übertragen werden. Somit bietet es sich an,

die Länge von S auf eine Sekunde festzulegen: Bei ausreichend genauen Messungen sind die Ergebnisse relativ aktuell. Geht man von einer Gesamtbandbreite von 11 Mbit/s aus, ist 0.2 Sekunden ein geeigneter Wert für S.

4.3.2 Schätzung der Bandbreitenauslastung

Die nach dem oben beschriebenen Point Sample Verfahren aufgenommenen Messwerte haben den Nachteil, dass einzelne Werte nach oben oder unten aus der Messreihe herausfallen können. Das bedeutet nicht notwendigerweise, dass sich wirklich Veränderungen in der Auslastung ergeben haben. Vielmehr können solche Schwankungen beispielsweise durch zufällige Pakethäufungen zustande kommen. Um eine Verfälschung der Messungen aufgrund solcher Ereignisse zu vermeiden, werden die Messwerte geglättet, indem sie zu bisherigen Werten ins Verhältnis gesetzt werden.

Auf diese Weise kann eine realistische Schätzung der Bandbreitenschätzung berechnet werden: In Anlehnung an das *Flip-Flop-Verfahren* [KN01] wird der Messwert *M* an zwei voneinander unabhängige *Exponentional Weighted Moving Average (EWMA)* Filter gegeben, die eine Verknüpfung mit vergangenen Werten herstellen. So entstehen zwei Schätzungen der aktuellen Netzauslastung (Abbildung 4.4 , (1)). Der eine Filter verwendet ein Gewicht von 0.9. Somit fließen neue Messwerte zu einem hohen Anteil ein. Der Filter hat nur eine geringe Trägheit und liefert somit eine agile Schätzung S_{agil}, glättet die Eingabewerte allerdings auch weniger stark:

$$S_{agil} = (1 - 0.9)S_{agil,alt} + 0.9M \ .$$

Im zweiten Filter kommt hingegen ein Gewicht von 0.1 zum Einsatz:

$$S_{stabil} = (1 - 0.1)S_{stabil,alt} + 0.1M \ .$$

Die Schätzungen S_{stabil} dieses Filters sind sehr stabil, Schwankungen in Messwerten werden gut geglättet. Das kleine Gewicht hat jedoch den Nachteil, dass es relativ lange dauert, bis der Schätzwert grundsätzliche Veränderungen der Netzlast widerspiegelt: Der Filter reagiert träge.

Insgesamt werden also zunächst zwei unterschiedliche Netzlastschätzungen ermittelt, die auf zwei unabhängig voneinander agierende EWMA-Filter zurückgehen. Sie sollen eine genau gegensätzliche Charakteristik haben. Dementsprechend und wegen der guten Ergebnisse in [KN01] wurden die beiden Gewichte wie dort extrem gewählt. Mit $0 < 0.1 < 0.9 < 1$ nutzen sie den möglichen Spielraum jeweils zu 90% aus.

Die weiteren Schritte (2) und (3) in Abbildung 4.6 dienen der Berücksichtigung von beginnenden und endenden Datenströmen und werden in Abschnitt 4.3.3 erläutert. Die Auswahl eines der beiden Filterwerte ((4) – (8)) wird in Abschnitt 4.3.4 beschrieben.

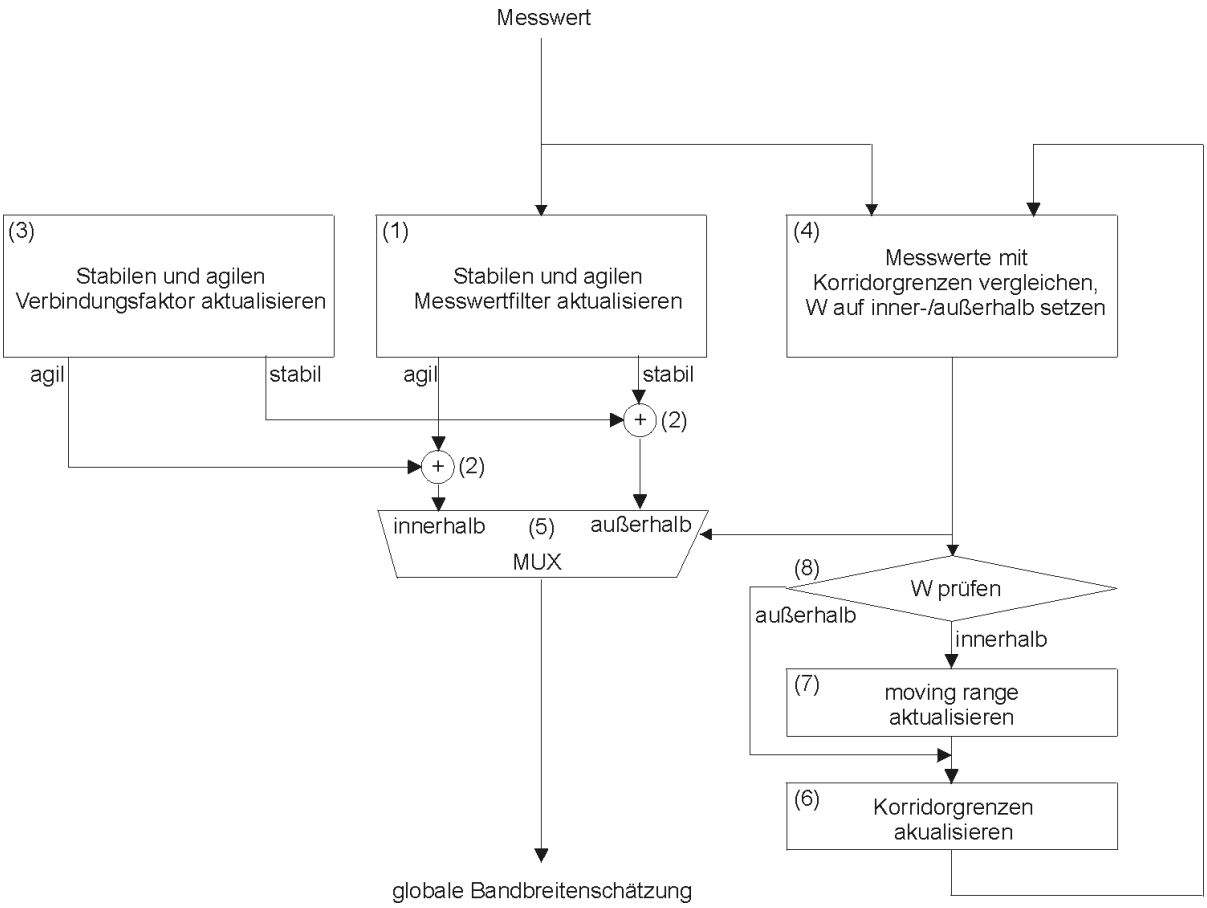

Abbildung 4.4: Berechnung der Bandbreitenschätzung aus den Messwerten

4.3.3 Berücksichtigung beginnender und endender Verbindungen

Große und schnelle Schwankungen der Zellauslastung entstehen, wenn neue Verbindungen zugelassen werden oder Verbindungen enden, wie im Beispiel zu Beginn der beiden eingezeichneten Intervalle (siehe *Bandbreite real* in Abbildung 4.5). Es ist wichtig, diese Schwankungen schnell zu erkennen, damit es nicht zu einem Unterschätzen (T1) oder Überschätzen (T2) der Netzauslastung kommt. Dies könnte beispielsweise zu falschen Entscheidungen der Zugangskontrolle führen.

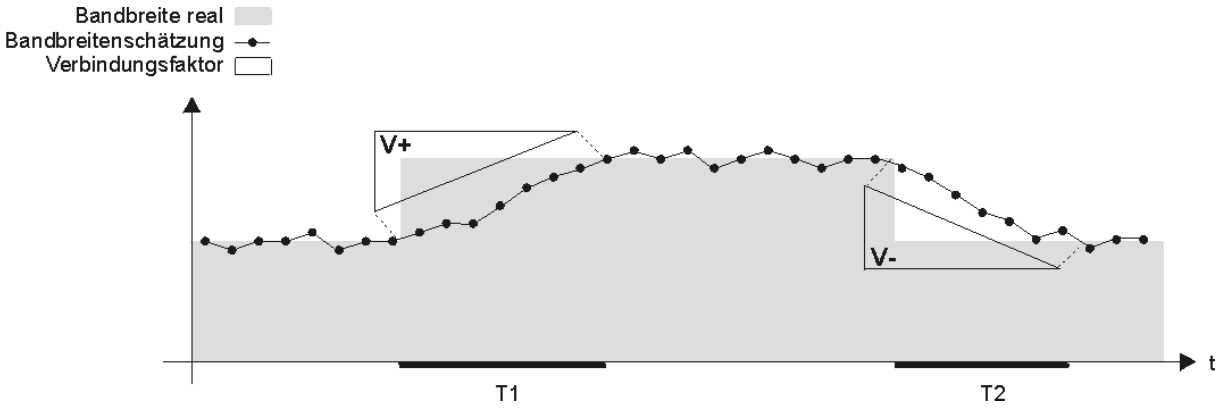

Abbildung 4.5: Einfluss von Verbindungsfluktuationen auf die Bandbreitenschätzung

EWMA-Filter haben den prinzipbedingten Nachteil, dass sie größere Veränderungen in der Netzauslastung (zum Beispiel durch eine neue Verbindung zu Beginn von T1) erst nach einer gewissen Zeit wiedergeben (siehe *Bandbreitenschätzung*). Die Schätzung liegt im Intervall T1 unterhalb der tatsächlichen Bandbreite. Dies gilt insbesondere für träge Filter mit einem kleinen Gewicht, grundsätzlich aber auch für agile Filter. Um Lastveränderungen durch Verbindungsfluktuation sofort in die Lastschätzung einbeziehen zu können, wird ein *Verbindungsfaktor V* eingeführt. Er wird regelmäßig aktualisiert und zu der aus den Messwerten generierten Schätzung hinzuaddiert (Abbildung 4.4, (2)). Er repräsentiert den noch nicht von den Filtern wiedergegebenen Bandbreitenanteil der neuen Verbindung (*V*+ in Abbildung 4.5). Wird eine neue Verbindung zugelassen, wird die von der Anwendung gegenüber der Zugangskontrolle angegebene Bandbreite zum Verbindungsfaktor hinzuaddiert. Bei Verbindungsende wird der betreffende Betrag abgezogen. Somit kann der Verbindungsfaktor auch negative Werte annehmen und so die Schätzung reduzieren (*V*-).

Werden die beiden Schätzungen aktualisiert, wird auch der Verbindungsfaktor angepasst (Abbildung 4.4, (3)): Da die Messungen Veränderungen in der Netzlast durch eine neue Verbindung nach und nach immer besser wiedergeben, wird der Faktor durch Multiplikation mit einem Gewicht < 1 immer mehr 0 angenähert:

$$V_{agil} = 0.1 \cdot V_{agil,alt} \quad V_{stabil} = 0.9 \cdot V_{stabil,alt} .$$

Damit der Verbindungsfaktor zu beiden Filtern passt, gibt es ihn in zwei Varianten mit unterschiedlichen Gewichten. Diese sind so gewählt, dass der Betrag des Verbindungsfaktors in dem Maß abnimmt, wie die Repräsentation der neuen Verbindung in den messungsbasierten Schätzungen zunimmt.

Die Einführung der Verbindungsfaktoren hat den Vorteil, dass durch beginnende oder endende Verbindungen hervorgerufene Auslastungsveränderungen sofort in der Summe aus Faktor und Schätzung repräsentiert werden. Auf diese Weise kann verhindert werden, dass die Auslastung über eine längere Zeit hinweg über- oder unterschätzt wird. Das Verfahren hat

allerdings den Nachteil, dass der anfängliche Betrag des Verbindungsfaktors von der a priori-Bandbreitenangabe der mobilen Anwendung abhängt.

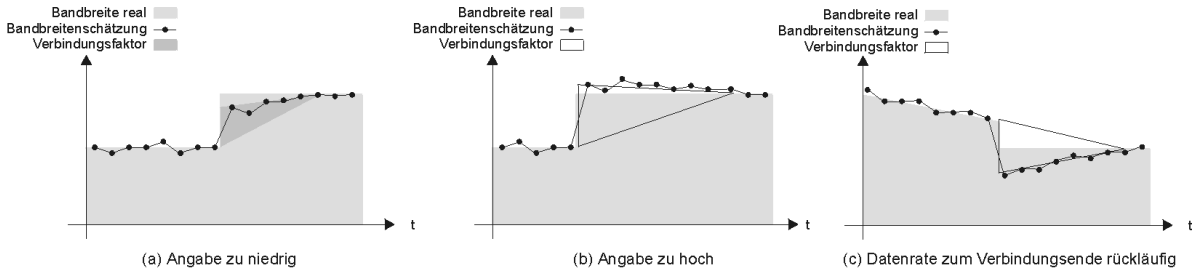

Abbildung 4.6: Auswirkungen falscher a priori-Angaben

Für neue Verbindungen ist das unproblematisch: Ist die Angabe zu niedrig, ist das Verfahren zwar weiterhin zu optimistisch, aber immerhin besser als ganz ohne Faktor (Abbildung 4.6a). Ist die Angabe zu hoch, ist man etwas zu konservativ (Abbildung 4.6b). Eine Verschlechterung gegenüber der Situation ohne Verbindungsfaktor tritt jedoch in einem speziellen Fall ein: Hat die mittlere Datenrate einer Verbindung zu ihrem Ende hin langsam abgenommen, haben sich die Schätzungen diesem Vorgang bereits angepasst (Abbildung 4.6c). Wird nun der Verbindungsfaktor in der ursprünglichen Höhe abgezogen, wird die Auslastung unterschätzt.

4.3.4 Auswahl eines der beiden Schätzwerte

Nach den beiden Schritten Schätzung und Verbindungsfaktor liegen nun zwei verschiedene Schätzungen der Netzwerklast vor, die beide begonnene und beendete Verbindungen berücksichtigen. Welche der beiden als Bandbreitenschätzung nach außen gereicht wird (externe Bandbreitenschätzung), wird durch einen Steuerkorridor geregelt (Abbildung 4.4, (4)). Liegen die einzelnen Messwerte außerhalb dieses Korridors, wird die stabile Schätzung verwendet, ansonsten die agile (5). Auf diese Weise werden die Messwerte genau dann stärker geglättet (und die Schätzungen träger), wenn die Messwerte stark schwanken. Sind sich die Messwerte relativ ähnlich, kann auf eine starke Glättung zu Gunsten einer geringeren Trägheit verzichtet werden.

Der Korridor wird oben und unten durch so genannte Steuergrenzen *G* beschränkt, die von der Varianz σ_M der Messwerte und einem Erwartungswert E abhängen. Wie in [KN01] sollen die Messwerte *M* als stark schwankend angesehen werden (und daher der stabile Filter zum Einsatz kommen), wenn sie stärker als $3\sigma_M$ vom Erwartungswert abweichen. Dies nennt man die 3-sigma-Regel. Als Erwartungswert E wird hier der agile oder stabile Schätzwert herangezogen, abhängig davon, ob der Messwert innerhalb oder außerhalb der letzten Korridorgrenzen lag. Insgesamt berechnen sich die neuen Korridorgrenzen (6) dann zu:

$$G = E \pm 3\frac{\overline{MR}}{d_2}.$$

Die *Moving Range* $\overline{MR}$ repräsentiert die Schwankungen der Messwerte: Zunächst wird die Differenz zwischen den letzten beiden Messwerten $|M_i - M_{i-1}|$ berechnet. Der Wert d_2 schätzt die Standardabweichung; wird $\overline{MR}$ aus zwei Messwerten ermittelt, nimmt d_2 etwa den Wert 1,128 an [KN01]. Um auch ältere Werte in die Varianzschätzung einzubeziehen, werden diese mittels eines EWMA-Filters gewichtet in $\overline{MR}$ einbezogen (7):

$$\overline{MR} = \frac{\overline{MR}_{alt}}{2} + \frac{|M_i - M_{i-1}|}{2}$$

Die Autoren in [KN01] geben nicht an, wie die beiden Summanden gegeneinander gewichtet werden sollen. Daher wurde hier eine 1:1-Gewichtung gewählt, um weder dem alten noch den neuen Wert eine dominierende Bedeutung zuzuordnen. $\overline{MR}$ wird nur dann aktualisiert, wenn der Messwert innerhalb des Korridors liegt (8).

Nach jedem Messvorgang werden somit einige Berechnungen durchgeführt, ihr Aufwand ist konstant.

4.3.5 Ein Rechenbeispiel für die Bandbreitenschätzung

Eine einfache Beispielrechnung (Abbildung 4.7) zeigt das Funktionsprinzip der Lastschätzung auf. Bei den Startwerten wurde davon ausgegangen, dass *M(t)* im Vorfeld so um 3000 schwankte, dass sich für $\overline{MR}$ ein Wert von 250 ergibt.

Bis zum Zeitpunkt t=4 liegen die Messwerte *M(t)* innerhalb der Korridorgrenzen. Somit wird als externe Schätzung S_{extern} die Summe aus S_{agil} und V_{agil} geliefert. Zu Beginn der fünften Messperiode wird eine neue Verbindung mit einer Datenrate von 3000 Byte/s zugelassen. Die Werte für den agilen und stabilen Verbindungsfaktor werden auf 3000 gesetzt und im weiteren Verlauf unterschiedlich schnell gegen Null geführt. Zum Ende der Periode 5 haben Verbindungsfaktoren dann noch die Werte 300 bzw. 2700. Der höhere Messwert liegt oberhalb des Korridors der vierten Periode. Daher wird $\overline{MR}$ nicht aktualisiert, und es wird auf die stabilen (trägeren) Varianten von Schätzung und Verbindungsfaktor umgeschaltet. Aufgrund des Verbindungsfaktors wird sich die neue Verbindung sofort in der externen Schätzung sichtbar. Zum Zeitpunkt t=6 wird wieder in den agilen Modus zurückgeschaltet.

t	$M(t)$	S_{agil}	S_{stabil}	V_{agil}	V_{stabil}	$\overline{MR}$	G_{unten}	G_{oben}	S_{extern}	
	3000	3000	3000	0	0	250	2500	3500	3000	Startwerte
1	3000	3000	3000	0	0	125	2667	3332	3000	
2	2800	2820	2980	0	0	163	2187	3053	2820	
3	3000	2982	2982	0	0	182	2498	3466	2982	
4	3100	3088	2994	0	0	141	2713	3463	3088	
5	6000	5708	3294	300	2700	141	5619	6369	5994	Neuzulassung (3000Byte/s)
6	6200	6151	3585	30	2430	170	5729	6633	6181	
7	5900	5925	3816	3	2187	235	2304	6553	5928	
8	6000	5992	4034	0	1968	268	5280	6705	5992	
9	8000	7799	4431	0	1771	268	5490	6914	6202	
10	6000	6179	4588	0	1594	1134	3164	9094	6179	
11	6000	6018	4729	0	1435	567	4511	7526	6018	

Abbildung 4.7: Beispielrechnung der Lastschätzung

Der Messwert der neunten Periode sticht deutlich nach oben heraus. Wieder wird der Korridor nach oben verlassen und daher in den stabilen Modus geschaltet. Es ist deutlich zu sehen, wie dadurch die Auswirkungen des Ausreißers auf die externe Schätzung begrenzt werden. Zum Zeitpunkt t=11 kann dann wieder der agile Filter verwendet werden.

4.4 Ressourcenreservierung

Die Aufgabe der Ressourcenreservierung ist erstens, die Höhe des Betrages der Reservierung in den Nachbarzellen zu ermitteln, welche sich aus den zur Zeit belegten Ressourcen ergibt. Diese muss den betreffenden Basisstationen mitgeteilt werden. Zweitens müssen Reservierungsnachrichten benachbarter Basisstationen angenommen und die betreffenden Reservierungen vorgenommen werden. Die Reservierungen werden dann von der Zugangskontrolle in die Zulassungsentscheidungen einbezogen.

Die Berechnung des Reservierungsbetrages für die Nachbarzellen geschieht auf Basis der Wechselwahrscheinlichkeiten und der aktuellen Ressourcenbelegung. Der Betrag ist virtuell. Das bedeutet, dass nicht in jeder Nachbarzelle die gesamten momentan belegten Ressourcen reserviert werden, sondern überall nur der Anteil, der der Wechselwahrscheinlichkeit dorthin entspricht. Die Reservierungen erfolgen aggregiert. Es werden somit nicht für einzelne Verbindungen Reservierungen vorgenommen, sondern die Gesamtheit der belegten

Ressourcen betrachtet. Durch die Einbeziehung der aktuell belegten Bandbreite erfolgt eine Adaption an das Kommunikationsverhalten der Netznutzer. Da sich die Reservierungshöhen an den Wechselwahrscheinlichkeiten der Bewegungsvorhersage orientieren, können sie sich auch an das Bewegungsverhalten der Netznutzer anpassen. Daraus resultieren unterschiedliche Beträge für die verschiedenen Nachbarzellen, die Reservierungen erfolgen also nicht-uniform.

4.4.1 Berechnung der Reservierungshöhe

Für den Betrag der Reservierungen gilt:

$$Reservierungen\ in\ Zelle\ i = p(i)\ S_{extern},$$

das heißt, die Reservierungshöhe für die Nachbarzelle i ergibt sich aus dem Produkt der Wechselwahrscheinlichkeit in diese Zelle und der Schätzung der Netzwerklast. Jede Basisstation sendet in Abständen von 1 Sekunde Reservierungsnachrichten an alle Nachbarzellen. Somit korrespondiert dieses Intervall mit dem Messintervall S: die Nachbarstationen sind jeweils auf einem aktuellen Stand, ohne dass zwei Pakete pro Sekunde (Nachricht und Antwort) einen nennenswerten Overhead erzeugen.

Der zugehörige Algorithmus lässt sich schematisch folgendermaßen skizzieren:

```
1.  currentUsedBandwidth = S_extern;
2.  for each neighbour cell i do
2.1  changeProbability = p(i);
2.2  reservation[i] = currentUsedBandwith * changeProbability;
2.4 od
3.  for each neighbour cell i do
3.1  “send reservation[i] to cell i“;
3.2 od
```

Sichtbar wird, dass die Berechnung von der Anzahl der Mobilfunkteilnehmer vollkommen unabhängig ist. Die Anzahl der Schleifendurchläufe hängt lediglich von der Zahl der benachbarten Basisstationen ab. Diese ist jedoch konstant und niedrig. Insbesondere aufgrund der relativ großen Periode können der Berechnungsaufwand und die daraus resultierende Signalisierung vernachlässigt werden.

Erhält eine Basisstation eine Reservierungsnachricht aus einer benachbarten Zelle, so löscht sie die alte Reservierung R_i der zugehörigen Basisstation, und speichert stattdessen die neue Reservierungshöhe ab. Daraufhin wird die Gesamthöhe R aller Reservierungen neu berechnet. Dieser Betrag steht nun nicht mehr für neu beginnende Verbindungen zur Verfügung. Anzumerken ist, dass die Summe aus belegten und reservierten Netzwerkressourcen die Gesamtressourcen überschreiten kann. Ein Beispiel für eine solche Situation wird in Kapitel 4.5 gegeben.

Als Antwort auf eine Reservierungsnachricht sendet die Basisstation eine Nachricht zurück, in der sie der reservierenden Station den Betrag der noch verfügbaren Bandbreite *F* mitteilt. Diese wird nach der Formel

$$F = B - \sum_{i=1}^{\#Nachbarn} R_i - S_{extern} = B - R - S_{extern}$$

berechnet. Die freie Bandbreite ergibt sich also aus der Gesamtbandbreite *B* abzüglich der Reservierungen und der aktuellen Auslastungsschätzung der Zelle.

Erhält eine Basisstation ein solches Antwortpaket, speichert sie die freie Bandbreite F_i dieser Basisstation ab. Daher weiß eine Basisstation immer, wie viel Bandbreite die umliegenden Stationen noch zur Verfügung haben. Diese Information kann dann von der verteilten Zugangskontrolle für die Zulassungsentscheidung genutzt werden.

4.4.2 Steuerparameter Reservierungshöhe

Im vorangegangenen Abschnitt wurde aufgezeigt, wie die Reservierungshöhe von HoPVarB aus dem Verhalten der Kommunikationsteilnehmer berechnet wird. Es ist jedoch unter Umständen aus Sicht des Netzbetreibers wünschenswert, auf die Höhe der Reservierungen manuell Einfluss nehmen zu können. Dies gilt insbesondere deshalb, weil reservierte Bandbreite natürlich nicht für neu beginnende Verbindungen genutzt werden kann. Daher soll in Anlehnung an [R01] ein Steuerparameter zur prozentualen Veränderung der Reservierungshöhe eingeführt werden. Er soll ermöglichen, dass unterschiedliche Reservierungs- und Auslastungsniveaus angestrebt werden können.

Die Steuermöglichkeit der *Reservierungshöhe* ergibt sich durch einen Faktor *H*, der in die Berechnung des Reservierungsbetrages einbezogen wird:

$$\textit{Reservierungen in Zelle } i = \frac{H}{1000}\, p(i)\, S_{extern}.$$

Dieser Steuerparameter kann Werte rund um 1000 annehmen und repräsentiert den Promilleanteil der vorgenommenen Reservierungen: Stellt man ihn beispielweise auf 900, so bewirkt dieses, dass von der berechneten Reservierungshöhe 10% abgezogen werden, bevor sie der Nachbarzelle mitgeteilt wird. Dabei kann *H* nicht nur Werte < 1000 annehmen; es ist auch möglich, den Reservierungsbetrag zu erhöhen.

Die Festlegung der Reservierungshöhe gilt global, das heißt im gesamten Funknetz für alle Zellen gleichermaßen.

Bei der Einstellung des Parameters auf Werte < 1000 sollte jedoch mit Bedacht vorgegangen werden. Eine zu starke Reduzierung der Reservierungen wird zwar die Auslastung verbessern, dieses allerdings auf Kosten einer steigenden Wahrscheinlichkeit von Handoverdrops.

4.5 Zugangskontrolle

Die Zugangskontrolle hat die Aufgabe, über die Zulassung von neu beginnenden und Handoververbindungen in einer Funkzelle zu entscheiden. Diese Entscheidung stützt sich seitens der Basisstation auf Informationen über belegte und reservierte Bandbreite. Sie soll einerseits sicherstellen, dass möglichst alle Verbindungen zugelassen werden können, die aufgrund eines Handovers in die Zelle kommen. Andererseits ist zu verhindern, dass insgesamt zu viele Verbindungen zugelassen werden, und so an bestehende Ströme gemachte Zusagen bezüglich der Bandbreite nicht eingehalten werden können.

Die Zulassungsentscheidung stützt sich auf eine Verbindungsanfrage seitens des Mobilfunkteilnehmers. Bei der Zugangskontrolle wird davon ausgegangen, dass die Angabe der von der mobilen Anwendung beantragten Datenrate prinzipiell korrekt ist. HoPVarB schließt keine Kontrolle des gesendeten Datenvolumens einzelner Verbindungen ein, diese Aufgabe muss von einer externen *Traffic Control* übernommen werden. Eine solche kann zum Beispiel auf einem Base Station Controller angeordnet sein.

Es ist die Aufgabe einer Handoverpriorisierung, Handovers gegenüber der Zulassung von neuen Verbindungen zu priorisieren. Somit unterscheiden sich die Zulassungskriterien für New Calls und Handovers naturgemäß.

4.5.1 Zugangskontrolle für New Calls

Um die Zulassung einer neuen Verbindung zu beantragen, sendet der Mobilfunkteilnehmer eine Nachricht an die Basisstation. Diese enthält die mittlere Bandbreite, die die Verbindung belegen wird.

Ein New Call wird zugelassen, wenn folgende Bedingungen erfüllt sind:

1. Die lokale Zelle hat genügend freie Bandbreitenressourcen zur Verfügung, um die neue Verbindung zu unterstützen, es muss gelten:

   $$B - R - S_{extern} \geq D.$$

   Das bedeutet, dass die Gesamtbandbreite der Basisstation abzüglich der Summe der Reservierungen und der externen Schätzung der aktuell belegten Bandbreite mindestens so groß sein muss wie die beantragte mittlere Datenrate *D*.

2. Alle benachbarten Basisstationen müssen genügend Bandbreitenressourcen zur Verfügung haben, um die aus einer Zulassung entstehenden Reservierungen vorzunehmen. Daher muss für alle benachbarten Funkzellen i = 1, ..., n gelten:

   $$F_i \geq \frac{H}{1000} w(i) D\,.$$

Das heißt, dass in allen benachbarten Funkzellen die freie Bandbreite mindestens so groß sein muss wie das Produkt aus Reservierungshöhe *H*, der beantragten mittleren Datenrate *D* und der Wechselwahrscheinlichkeit in die betreffende Zelle.

Die erste Bedingung macht deutlich, dass die Zugangskontrolle messungsbasiert arbeitet: In die Berechnung der freien Bandbreite *F* fließt die Netzlastschätzung ein. In der zweiten Bedingung ist erkennbar, dass die Zugangkontrolle für New Calls verteilt erfolgt, die umliegenden Basisstationen werden in die Entscheidung einbezogen. Sie basiert zudem auf der mittleren Datenrate. Eine Diskriminierung von Strömen mit hoher Peakrate und niedriger mittlerer Datenrate kann somit ausgeschlossen werden.

Eine Zulassung einer neuen Verbindung erfolgt nur, wenn beide Bedingungen erfüllt sind. Anderenfalls kommt es zu einem New Call Block.

4.5.2 Zugangskontrolle für Handover Calls

Die Zulassungsentscheidung für Verbindungen, die aufgrund eines Zellenwechsels beantragt werden, wird anhand anderer Kriterien gefällt. Ein Handover Call wird zugelassen, wenn für seine mittlere Datenrate gilt:

$$B - S_{extern} \geq D.$$

Eine Zulassung erfolgt also, wenn die Gesamtbandbreite abzüglich der zur Zeit durch Datenverkehr belegten Bandbreite mindestens so groß ist wie die mittlere Datenrate *D*. Die reservierte Bandbreite *R* wird nicht berücksichtigt, da sie für Handoververbindungen vorgesehen ist.

Daher kann es zeitweilig dazu kommen, dass die Summe aus Reservierungen und belegter Bandbreite die Gesamtbandbreite übersteigt, beispielsweise nach einer Serie von Zellwechseln. Dies wird dazu führen, dass keine neuen Verbindungen mehr in dieser Zelle zugelassen werden; somit wird die Zahl der aktiven Verbindungen zurückgehen, wenn Verbindungen enden. Handover Calls hingegen werden weiterhin zugelassen, solange die Gesamtbandbreite unterhalb der Schätzung der Zellauslastung liegt.

Im Falle einer positiven Zulassungsentscheidung, egal ob bezüglich eines New Calls oder eines Handover Calls, wird die beantragte Bandbreite der neuen Verbindung der Komponente zur Messung der Zellauslastung mitgeteilt. Diese kann dann die Verbindungsfaktoren V_{agil} und V_{stabil} durch hinzuaddieren von *D* aktualisieren.

4.6 Eigenschaften

Im Folgenden wird aufgezeigt, dass HoPVarB den in Kapitel 3.1 aufgestellten Anforderungen genügt.

4.6.1 Effizienz

Die wichtigste Anforderung ist eine gute Netzwerkauslastung. Dem wird Rechnung getragen, indem durch Messung der Zellauslastung näherungsweise nur der Betrag an Bandbreite als belegt angesehen wird, den die Datenströme insgesamt zur Datenübertragung nutzen. Im Gegensatz zu einem Peakrate-Ansatz wird weniger Bandbreite belegt. Durch das Verfahren zur Schätzung der Zellauslastung wird sichergestellt, dass sich Veränderungen der Netzlast durch beginnende oder endende Verbindungen sofort in den Schätzungen niederschlagen. Das Schätzmodul arbeitet effektiv, Schwankungen der Messwerte nach oben (oder unten) führen nicht zu überpessimistischen (oder zu optimistischen) Schätzungen. Durch die Messungsbasierung arbeitet die Zugangsgangskontrolle somit äußerst effektiv.

Auch die Ressourcenreservierung beruht auf den Messungen und reserviert im Gegensatz zu Peakrate-Ansätzen nur die Bandbreite, die im Falle eines Handover im Mittelwert benötigt wird. Da die Reservierungen virtuell erfolgen, wird der Reservierungsbetrag im Vergleich zu einer nicht-virtuellen Vorgehensweise deutlich reduziert, ohne dass es im statistischen Mittel zu Unterreservierungen kommt. Da die Reservierungen sich mittels der Bewegungsvorhersage am Verhalten der Nutzer orientieren, werden diese deutlich effizienter vorgenommen als bei nicht-adaptiven, uniformen Verfahren.

4.6.2 Skalierbarkeit

Sowohl die Speicher- als auch die Berechnungskomplexität des Verfahrens ist von Nutzerzahl und Netzgröße unabhängig. Lediglich die Anzahl der Nachbarzellen hat hier Einfluss, diese kann jedoch als kleine Konstante angesehen werden. Messungen und Reservierungen werden aggregiert durchgeführt. Die Datenhaltung erfolgt dezentral auf den Basisstationen, Kommunikation fällt nur mit den direkt benachbarten Zellen an. Sie wird periodisch und unabhängig von beginnenden Verbindungen durchgeführt.

4.6.3 Einfachheit und Berechenbarkeit

Das Verfahren ist einfach und transparent. Es gibt vier Komponenten mit eindeutig abgegrenzten Aufgaben. Der einzige externe Steuerparameter hat eine klar absehbare Wirkungsweise. Die einfachen Berechnungen sind lediglich aus den Grundrechenarten konstruiert und der Aufwand des Verfahrens kann als konstant angesehen werden. Daher kann davon ausgegangen werden, dass HoPVarB sich auch in einem Umfeld mit bescheidener Hardwareausstattung einsetzen lässt. Außerdem basiert der Algorithmus nicht auf komplexen theoretischen Annahmen, die das Verständnis erschweren.

4.6.4 Robustheit

Durch die beiden im Verfahren zur Schätzung der Netzwerkauslastung eingesetzten Filter ist HoPVarB äußerst robust gegen schwankende Messwerte. Dennoch werden Lastveränderungen, die durch beginnende oder endende Verbindungen hervorgerufen werden, sofort in den Schätzungen wiedergegeben. Dies kann allerdings in dem Spezialfall, dass die Datenrate einer Verbindung zum Ende hin abnimmt, zu einer kurzfristigen Unterschätzung der Auslastung führen.

Da sich geringe Abweichungen zwischen beantragter und tatsächlicher mittlerer Datenrate nicht auf spätere Zulassungsentscheidungen der Zugangskontrolle auswirken, ist die Fehlertoleranz diesbezüglich gut. Auch können sich die Datenstromeigenschaften einer Verbindung nach der Zulassung durchaus geringfügig ändern, ohne dass es zu einer Neuzulassung kommen muss. Das liegt daran, dass diese Veränderungen durch die kontinuierlichen Messungen von HoPVarB bemerkt werden, und so in zukünftige Zulassungsentscheidungen einbezogen werden können. Dennoch ist offenbar, dass insbesondere ein schnelles Zunehmen der mittleren Datenrate einer Verbindung nicht hinnehmbar ist, da es sonst zu Überlastsituationen innerhalb einer Zelle kommen kann. Solche Veränderungen müssen von einer Traffic Control reglementiert werden.

Das periodische Versenden von Reservierungs- und Registrierungsnachrichten hat den Vorteil, dass auch Paketverluste oder ausfallende Mobilteilnehmer nicht zu Verklemmungen oder nicht mehr frei werdenden Ressourcen führen.

4.6.5 Fairness

Die Zugangskontrolle stützt sich auf die mittlere Datenrate der Verbindungen anstatt auf ihre Peakrate. Daher ist der Netzzugang ohne Benachteiligung auch für extrem burst-artige Verbindungen möglich. Allerdings ist es bei zunehmender Netzauslastung weiterhin so, dass Ströme mit einer niedrigen Datenrate eher zugelassen werden als solche, die viele Ressourcen belegen und somit die Kapazität überschreiten würden. Dies liegt jedoch in der Natur der begrenzten Ressourcen und ließe sich lediglich verhindern, indem man verschiedene Klassen mit eigenen Ressourcen für Ströme unterschiedlicher Bandbreite definiert.

4.7 Zusammenfassung

Im diesem Kapitel wurde ein Algorithmus zur Handoverpriorisierung für Anwendungen mit variabler Bitrate in mobilen DiffServ Netzen (HoPVarB) entwickelt. Er gliedert sich in die vier Komponenten Bewegungsvorhersage, Schätzung der Zellenauslastung, Ressourcenreservierung und Zugangskontrolle.

Die mittels einer dezentralen Datenbasis umgesetzte historienbasierte Bewegungsvorhersage sorgt für gute Adaptivität bei gleichzeitig guter Skalierbarkeit. Die auf aggregierten Messungen basierende Komponente zur Schätzung der Netzwerkauslastung arbeitet mit zwei verschiedenen EWMA-Filtern, die unabhängig voneinander arbeiten. Durch einen Steuerkorridor wird abhängig von der Streuung der Messwerte der agilere oder der stabilere der beiden Filter gewählt. Zusätzlich werden Informationen über beginnende und endende Verbindungen einbezogen, um zu einer realistischen Schätzung zu gelangen. Die Ressourcenreservierung arbeitet aggregiert, virtuell und adaptiv. Sie tauscht periodisch Nachrichten mit den Nachbarzellen aus. Dabei lässt sich die Reservierungshöhe durch einen Steuerparameter manuell beeinflussen. Die Zugangskontrolle arbeitet verteilt, das heißt, sie bezieht die Ressourcenbelegung in den Nachbarzellen in die Zulassungsentscheidung ein. Sie stützt sich dabei auf die mittlere Datenrate des neuen Stromes.

Abschließend wurde aufgezeigt, dass HoPVarB den in Kapitel 3.1 aufgestellten Anforderungen Effizienz, Skalierbarkeit, Einfachheit, Robustheit und Berechenbarkeit genügt.

5 Simulation des Verfahrens

Um die Funktionsfähigkeit des im vorigen Kapitel entwickelten Algorithmus zu überprüfen und seine Leistungsfähigkeit zu belegen, wurde das Verfahren in einer Vielzahl von Netzwerksimulationen erprobt und bewertet.

Das vorliegende Kapitel beschreibt diese Simulationen und ihre Ergebnisse. In einem ersten Abschnitt werden grundlegende Eigenschaften des Simulationsszenarios beschrieben. Danach wird der verwendete Simulator vorgestellt und eine Einführung in die Simulationsumgebung gegeben. Dabei wird auch die Integration der Handoverpriorisierung in Mobile IP dokumentiert. Im darauf folgenden Abschnitt werden die verwendeten Verkehrsmodelle erläutert. Abschließend werden die einzelnen Szenarien detailliert beschrieben und die Simulationsergebnisse dokumentiert.

5.1 Grundlagen der Simulation mobilen Netzwerkverkehrs

Im Vergleich zur Simulation drahtgebundener Netzwerke erfordert die Simulation mobilen Netzwerkverkehrs einen gewissen Mehraufwand bei der Definition der Szenarien. In drahtgebundenen Szenarien müssen lediglich die Hosts mit ihren Datenquellen und -senken sowie die Verbindungen zwischen ihnen (*Topology*) festgelegt werden. Bei Mobilszenarien kommen die Definition von Anordnung und Größe der Funkzellen (*Scenery*) sowie Position und Bewegung der Mobilfunkteilnehmer (*Mobility*) hinzu.

5.1.1 Netzwerktopologie

Abbildung 5.1 zeigt die für die Simulationen verwendete Topologie, das heißt die Netzwerkknoten sowie die Verbindungen (*Links*) zwischen ihnen.

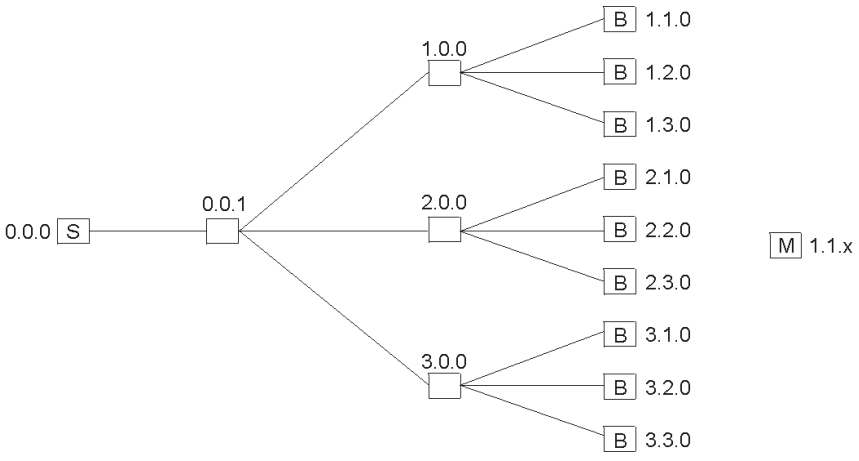

Abbildung 5.1: Topologie des simulierten Netzwerkes

Direkt an den abgebildeten Knoten sind ihre jeweiligen Adressen vermerkt. Sie sind spezifisch für den verwendeten Simulator und mit einer IP-Adresse vergleichbar. Es kommen neun Basisstationen (B) zum Einsatz, die jeweils eine Kapazität von X Bandbreiteneinheiten haben. Sowohl X als auch die Bandbreite der einzelnen Einheiten sind jeweils konfigurierbar. Je drei Basisstationen sind mit einem weiteren Knoten verbunden. Die Bandbreite der betreffenden Netzwerkverbindungen ist jeweils so gewählt, dass sie genau der Kapazität der Basisstation entspricht. Über einen weiteren Knoten sind die Stationen mit der Datensenke (S) verbunden, die den Netzwerkverkehr aufnimmt. Die betreffenden Links sind überdimensioniert, um keine unerwünschten Flaschenhälse entstehen zu lassen. Die Mobilfunkteilnehmer (M) dienen in den Simulationen als Datenquellen, sie generieren Datenverkehr. Obwohl in der Realität Quellen und Senken in der Regel eher umgekehrt angeordnet sind, wurde diese Konfiguration gewählt, um die Simulationen zu vereinfachen. So wird vermieden, dass der Mobilfunkteilnehmer den Datenstrom zunächst aus dem Netz anfordern muss. Auf die Aussagekraft der Simulationsergebnisse hat dieser Umstand keinen Einfluss, da die Richtung des Datenstromes für das simulierte Verfahren nicht von Bedeutung ist.

5.1.2 Scenery

Räumlich wurden die neun Basisstationen (schwarz gestrichelt) in einem drei-mal-drei-Raster angeordnet (Abbildung 5.2).

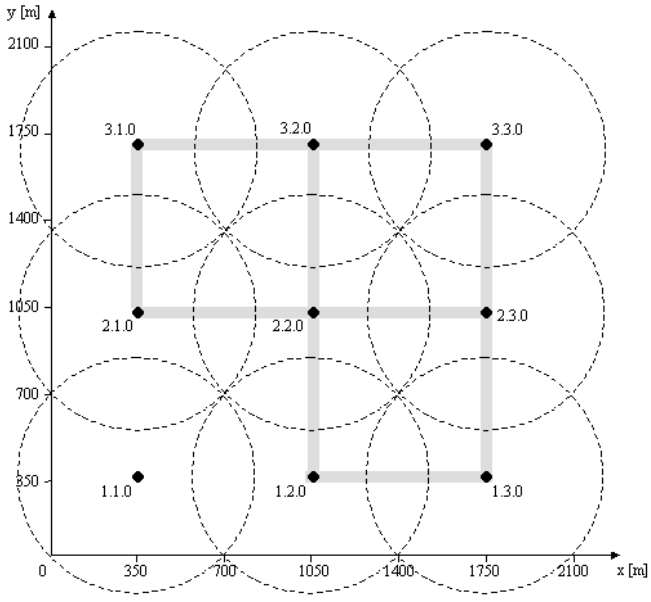

Abbildung 5.2: Anordnung der Funkzellen und Bewegungskorridore

Sie haben einen Abstand von 700 Metern. Jede Funkzelle hat (sofern sie nicht am Rand steht) vier Nachbarzellen. Deren Anzahl ist jedoch prinzipiell beliebig und wurde so gewählt, dass

die grau eingezeichneten Bewegungskorridore der Mobilfunkteilnehmer jeweils eindeutig von einer Zelle in eine andere verlaufen.

5.1.3 Bewegungsmodell

Die Bewegungskorridore resultieren aus einem Bewegungsmodell, das verschiedene Typen von Bewegungszellen verwendet (Abbildung 5.3). Der Verlauf der Bewegungskorridore entsteht durch die Anordnung der einzelnen Zellen. Er wurde so gewählt, dass in der mittleren Zelle (2.2.0) eine besonders starke Verkehrskonzentration entsteht, denn hier muss fast jeder Mobilfunkteilnehmer vorbei. In abgeschwächter Form kommt es auch in der mittleren oberen Zelle (3.2.0) und in der mittleren rechten Zelle (2.3.0) zu einem erhöhten Verkehrsaufkommen.

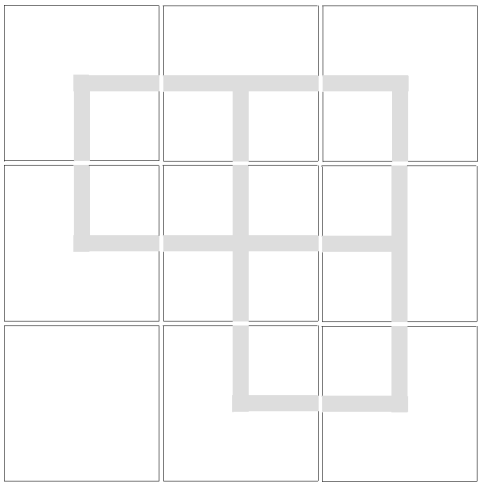

Abbildung 5.3: Aus den neun Bewegungszellen gebildeter Bewegungskorridor

Jede der Bewegungszellen ist quadratisch und hat einen bestimmten Straßenverlauf (grau). Aus diesem ergeben sich Richtungen, in die sich der Mobilfunkteilnehmer von der Zellmitte aus bewegen kann. Bei mehreren möglichen Bewegungsrichtungen sind die Wahrscheinlichkeiten jeweils gleich verteilt. So steuern die Mobilfunkteilnehmer jeweils die Mitte einer benachbarten Bewegungszelle an. Dort angekommen, wählen sie anhand der Richtungswahrscheinlichkeiten des betreffenden Zelltyps eine neue Zielzelle aus. Die Zelle, aus der der Teilnehmer gekommen ist, steht dabei als Ziel nicht zur Verfügung, die verbleibenden Richtungswahrscheinlichkeiten werden dementsprechend so skaliert, dass die Summe wiederum eins ergibt. Abbildung 5.4 gibt einen Überblick über die verwendeten Typen von Bewegungszellen mit den dazugehörigen Wechselwahrscheinlichkeiten.

Zelltyp	Wechselwahrscheinlichkeit in Richtung				Aussehen
	y	x	-y	-x	
Kreuzung	0.25	0.25	0.25	0.25	
Leerzelle	0	0	0	0	
T-Kreuzung	0.33	0	0.33	0.33	
	0	0.33	0.33	0.33	
Knick	0.5	0.5	0	0	
	0	0	0.5	0.5	
	0	0.5	0.5	0	
	0.5	0	0	0.5	

Abbildung 5.4: Verwendete Bewegungszelltypen mit Wechselwahrscheinlichkeiten

Für jede Zelle lässt sich darüber hinaus eine Maximalgeschwindigkeit festlegen, mit der die Mobilfunkteilnehmer sich höchstens bewegen können. Auf diese Weise lassen sich zum Beispiel Staus simulieren. In den Simulationen dieser Arbeit war die Höchstgeschwindigkeit für alle Zellen einheitlich auf 17 m/s eingestellt. Dies entspricht etwa den im Stadtverkehr üblichen 60 km/h. Alle Bewegungszellen haben eine Größe von 700 mal 700 Meter. Sie sind räumlich deckungsgleich mit den Funkzellen, d.h. die Basisstation steht jeweils genau in der Mitte der Bewegungszelle.

Definition der Netzwerklast

Von einem Mobilitätsgenerator wird eine bestimmten Menge an Mobilfunkteilnehmern simuliert, so dass eine vorher definierte Netzwerklast entsteht. Eine Last von 100% bedeutet dabei, dass die Summe der mittleren Bitraten der Verbindungswünsche, die dabei auf jede Basisstation entfallen, genau deren Bandbreitenkapazität entspricht.

Die Anzahl der Mobile Hosts ergibt sich aus der Anzahl der Zellen, der Netzlast pro Teilnehmer und der gewünschten Last pro Zelle. Die Mobilfunkteilnehmer werden gleichmäßig in allen Zellen mit Bewegungskorridor (also nicht in Leerzellen) verteilt. Ihre Bewegungs- und somit auch Verbindungszeiten sind um 180 Sekunden exponentialverteilt. Hat ein Teilnehmer seine Verbindung beendet, wird er neu positioniert und beginnt eine

weitere Verbindung. Dabei wird er so eingesetzt, dass die Vorgaben bezüglich der Interankunftsraten der Verbindungen und der Netzwerklast stets eingehalten werden.

Während einer Verbindung erzeugt ein Mobilfunkteilnehmer Datenverkehr, der einem bestimmten Verkehrsmodell entspricht. Diese Modelle werden in Abschnitt 5.4 detailliert vorgestellt.

5.2 Simulationsumgebung

5.2.1 Der Simulator ns-2

Die simulative Erprobung des Verfahren wurde mit dem Network Simulator 2 (ns-2) [NS2] durchgeführt. Bei ns-2 handelt es sich um einen ereignisgesteuerten Simulator, der für die Simulation von Netzwerkverkehr entwickelt wurde.

ns-2 unterstützt eine Vielzahl von Netzwerkprotokollen, darunter auch Protokolle zur Simulation von drahtlosem, mobilem Netzwerkverkehr. Der Simulator hat den Vorteil, dass er die Kommunikation im Netzwerk sehr genau nachempfindet. Alle beteiligten Protokollschichten werden realitätsnah und auf der Ebene der einzelnen Pakete simuliert. Darüber hinaus handelt es sich bei ns-2 um eine Public-Domain-Software, das heißt sie steht kostenlos zur Verfügung.

Die Beschreibung einer Simulation besteht aus einer statischen und einer dynamischen Komponente. Der statische Teil beschreibt die an der Simulation beteiligten Objekte, ihre Eigenschaften und ihre Beziehungen zueinander. Dazu gehören beispielsweise die Netzwerktopologie oder Eigenschaften von Netzknoten. Der dynamische Teil hingegen beschreibt Aktionen dieser Objekte, beispielsweise das Versenden eines Paketes. Aktionen haben somit immer einen Zeitbezug und werden als Ereignisse bezeichnet. Ein Simulator ist *ereignisgesteuert*, wenn er das simulierte Geschehen in Form von Ereignissen abbildet und diese der Reihe nach abarbeitet. Dazu werden die Ereignisse chronologisch in eine Warteschlange eingereiht. Ein Scheduler verwaltet diese Warteschlange (Abbildung 5.5).

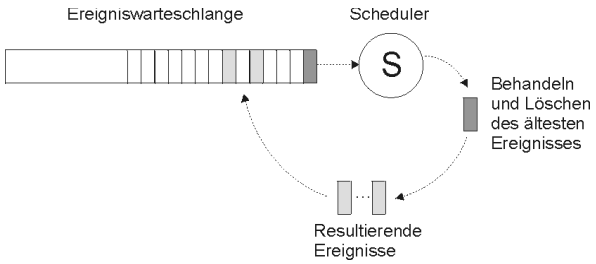

Abbildung 5.5: Funktionsweise eines ereignisgesteuerten Simulators

Erreicht die simulierte Zeit den Zeitpunkt des ältesten Ereignisses in der Warteschlange, wird dieses aus der Schlange entfernt. Bei dessen Behandlung können (mehrere) resultierende

Ereignisse entstehen. So kann beispielsweise das Versenden eines Paketes in einer Protokollschicht dessen Ankunft in der darunter liegenden Schicht auslösen. Der Zeitpunkt dieser Folgeereignisse wird errechnet, die entsprechenden Ereignisse werden erzeugt und an der richtigen Stelle in die Warteschlange eingereiht. Die simulierte Zeit wird während der Ereignisbehandlung angehalten, es wird also jeweils nur ein Ereignis gleichzeitig behandelt. Somit handelt es sich um eine diskrete Ereignissteuerung.

ns-2 ist ein objektorientierter Simulator, der bilingual implementiert ist [FV02]. Der Simulator ist in C++ [S00] geschrieben und hat einen OTcl [O94] Interpreter als Frontend. Er unterstützt eine Klassenhierarchie in C++ (die sogenannte *übersetzte Hierarchie*), und eine ähnliche Hierarchie innerhalb des OTcl Interpreters (die *interpretierte Hierarchie*). Beide Hierarchien sind eng miteinander verbunden, aus Sicht des Nutzers gibt es eine Art Eins-zu-Eins-Beziehung: Für nahezu jede OTcl-Klasse gibt es ein entsprechendes C++-Pendant und umgekehrt. Der Nutzer kann so Simulatorobjekte mittels des Interpreters erzeugen: Sie werden innerhalb des Interpreters instanziert, und eng von einem Objekt der übersetzten Hierarchie widergespiegelt, das automatisch erzeugt wird. ns-2 sieht spezielle Programmierkonstrukte vor, die die beiden Hierarchien miteinander verbinden und unter anderem die gemeinsame Nutzung von Variablen und den gegenseitigen Aufruf von Methoden ermöglichen.

ns-2 nutzt zwei Programmiersprachen, da der Simulator zwei unterschiedlichen Anforderungen gerecht werden muss: Einerseits erfordert die detaillierte Simulation von Protokollen und Paketen einen schnellen, effizienten Code. In diesem Bereich, der Implementierung von Protokollinstanzen (sogenannten *Agenten*), wird C++ eingesetzt. Andererseits muss die Konfiguration von Simulationen schnell und einfach möglich sein. Aufgrund der Tatsache, dass die Interpretersprache OTcl keine Neuübersetzung nach einzelnen Änderungen erfordert, ist sie hier besonders geeignet. Somit wird diese objektorientierte Variante von Tcl als Kommando- und Konfigurationssprache eingesetzt, ns-2 wird beim Start ein OTcl-Skript als Parameter übergeben, das der integrierte Interpreter abarbeitet.

5.2.2 Konfigurationsbeispiel

In Abbildung 5.6 ist beispielhaft ein Ausschnitt eines solchen Konfigurationsskriptes aus [TUT] zu sehen. Es zeigt, wie die Simulation einer einfachen Netzwerkkommunikation zweier Knoten konfiguriert und gestartet werden kann.

```
01 #Simulatorinstanz erzeugen
02 set ns [new Simulator]
03
04 #zwei Knoten instanzieren
05 set n0 [$ns node]
06 set n1 [$ns node]
07
16 #Datenverkehr konfigurieren
17 set cbr0 [new Application/Traffic/CBR]
18 $cbr0 set packetSize_ 500
19 $cbr0 set interval_ 0.005
20 $cbr0 attach-agent $udp0
21
22 #Verbindung definieren
```

```
08 #einen Link zwischen den Knoten        23 $ns connect $udp0 $null0
09 $ns duplex-link $n0 $n1 1Mb 10ms       24
                                  DropTail 25 #Simulationsereignisse festlegen
10                                        26 $ns at 0.5 "$cbr0 start"
11 #Protokollinstanzen erzeugen           27 $ns at 4.5 "$cbr0 stop"
12 set udp0 [new Agent/UDP] #Datenquelle  28 $ns at 5.0 "exit 0"
13 $ns attach-agent $n0 $udp0             29
14 set null0 [new Agent/Null] #Datensenke 30 #Simulator starten
15 $ns attach-agent $n1 $null0            31 $ns run
```

Abbildung 5.6: Einfaches Beispiel für ein OTcl Konfigurationsskript

Zunächst wird ein Simulatorobjekt instanziert (Zeile 2). Dann beginnt die statische Simulationsbeschreibung: Es werden die beiden Netzwerkknoten erzeugt (Zeilen 5 und 6) und zwischen ihnen eine physikalische Netzwerkverbindung mit einer Bandbreite von 1 Mbit/s und einer Verzögerung von 10 ms hergestellt (Zeile 9). In den Zeilen 12 bis 15 werden ein UDP-Agent und ein Datensenken-Agent an die Knoten angeschlossen. Dann wird ein CBR-Verkehrsgenerator erzeugt, der im Abstand von 0.005 Sekunden Pakete der Größe 500 Byte generiert, und mit dem UDP-Agenten verbunden (Zeilen 17 bis 21). Zeile 23 stellt eine UDP-Verbindung zwischen den Knoten her. Dann folgt die dynamische Simulationsbeschreibung (Zeilen 26 bis 28): Zum Zeitpunkt 0.5 Sekunden wird der Verkehrsgenerator gestartet, nach 4 weiteren Sekunden wird er wieder gestoppt. Zum Zeitpunkt 5 Sekunden soll ns-2 den Simulationslauf dann beenden. Abschließend wird der Simulator gestartet (Zeile 31).

5.2.3 Konfiguration von Simulationen mobilen Netzwerkverkehrs

Wie bereits in Kapitel 5.1 beschrieben, geht mit der Simulation von drahtlosem Netzwerkverkehr ein erhöhter Konfigurationsaufwand im Vergleich mit der Simulation drahtgebundener Kommunikation einher. Im Hinblick auf die statischen Aspekte genügt es nicht, lediglich die bereits vorgestellte Topologie zu konfigurieren, darüber hinaus müssen die ortsbezogenen und funknetztechnischen Eigenschaften der Basisstationen sowie die Konfiguration der Mobilfunkteilnehmer vorgenommen werden. Hinsichtlich der dynamischen Aspekte kommt die Platzierung und Bewegung der Mobilfunkteilnehmer hinzu. Diese Konfigurationsmaßnahmen sollen im Folgenden schematisch vorgestellt werden; für eine umfassende Beschreibung sei auf [FV02] verwiesen.

Die statische Konfiguration beginnt mit der Festlegung der Ausdehnung des Geländes:

```
set topo [new Topography]
$topo load_flatgrid x-Wert y-Wert
```

Hier wird ein rechteckiges, flaches Gebiet mit einer Ausdehnung von `x-Wert` mal `y-Wert` Metern erzeugt. Darüber hinaus muss ein General Operations Director (God) erzeugt werden:

```
set god_ [create-god Anzahl-Knoten]
```

Dieses Objekt speichert allgemein Informationen über den Zustand der Umwelt, des Netzes oder der Knoten, die ein allwissender Zuschauer hätte, die jedoch kein Simulationsteilnehmer speichern soll [TUT].

ns-2 unterstützt eine Vielzahl von unterschiedlichen Protokollen zur Simulation mobilen Netzwerkverkehrs. Bevor Knoten erzeugt werden können, die mit drahtlosem Verkehr allgemein bzw. mit Mobile IP im Speziellen befasst sind, ist somit eine Vielzahl von Einstellungen zu treffen, damit für die verschiedenen Schichten das jeweils richtige Protokoll ausgewählt wird:

```
$ns_ node-config -wiredRouting   ON
         –mobileIP               ON
         -adhocRouting           NOAH
         -macType                MAC/802_11
         -antType                Antenna/OmniAntenna
         -propType               Propagation/TwoRayGround
         -ifqType                Queue/DropTail
         -phyType                Phy/WirelessPhy
         -topoInstance           $topo
         -channel                $chan_1_
         :                       :
```

Besonders wichtig unter diesen Optionen ist der erste Eintrag. Für eine Basisstation muss `-wiredRouting` auf `ON` gestellt werden, für einen Mobilfunkteilnehmer hingegen auf `OFF`. Für beide Knotentypen wird die Verwendung von Mobile IP eingeschaltet. Weiterhin werden Routingverfahren, Medienzugangsverfahren, Antennentyp und vieles andere mehr festgelegt. Eine vollständige Dokumentation aller Optionen findet sich in [FV02].

Die Position der Knoten in dem flachen, rechteckigen Gelände `$topo` wird mit

```
$Knoten set X_ x-Wert
$Knoten set Y_ y-Wert
```

festgelegt. Bei den Mobilfunkteilnehmern muss zusätzlich der Home Agent festgelegt werden:

```
[$Knoten set regagent_] set home_agent_ Adresse
```

Die wesentlichen Elemente der dynamischen Konfiguration sind die Bewegungen sowie Beginn und Ende des Datenverkehrs der einzelnen Mobilfunkteilnehmer. Beides wird wie oben bereits beschrieben durch den Verkehrsgenerator automatisch für alle mobilen Knoten berechnet und in einer separaten Datei gespeichert, die in das Konfigurationsskript eingebunden wird. Sie besteht aus Sätzen von Einträgen wie:

```
$Knoten set X_ x-Wert
$Knoten set Y_ y-Wert
```

```
$ns at Zeitpunkt "[$Knoten set traffic_generator] clear"
$ns at Zeitpunkt "[$Knoten set regagent_] turnOn"
$ns at Zeitpunkt "$Knoten setdest x-Wert y-Wert Geschw"
:
$ns at Zeitpunkt "[$Knoten set traffic_generator] stop"
$ns at Zeitpunkt "[$Knoten set regagent_] turnOff"
```

Zunächst wird der Host positioniert, dann wird sein Verkehrsgenerator zurückgesetzt und der Mobile IP Agent eingeschaltet. Es folgt eine bestimmte Anzahl von Anweisungen, die den Teilnehmer jeweils mit einer bestimmten Geschwindigkeit auf einen Punkt zustreben lassen. Diese Anweisung gilt so lange, bis sie von einer anderen ersetzt wird. Abschließend werden Verkehrsgenerator und Agent abgeschaltet. Über die verschiedenen Verkehrsgeneratoren, die bei den Simulationen zum Einsatz gekommen sind, finden sich in Abschnitt 5.4 detaillierte Ausführungen.

5.3 Integration von HoPVarB in ns-2

Um die Funktions- und Leistungsfähigkeit von HoPVarB überprüfen zu können, musste das Verfahren in den Netzwerksimulator integriert werden. Hier wurde Mobile IP als Integrationspunkt gewählt. Das bedeutet aber nicht, dass nicht auch eine anderweitige Integration möglich wäre.

5.3.1 Integration der Komponenten

Die wesentlichen Komponenten von HoPVarB wurden im und um den Mobile-IP-Agenten angeordnet (Abbildung 5.7).

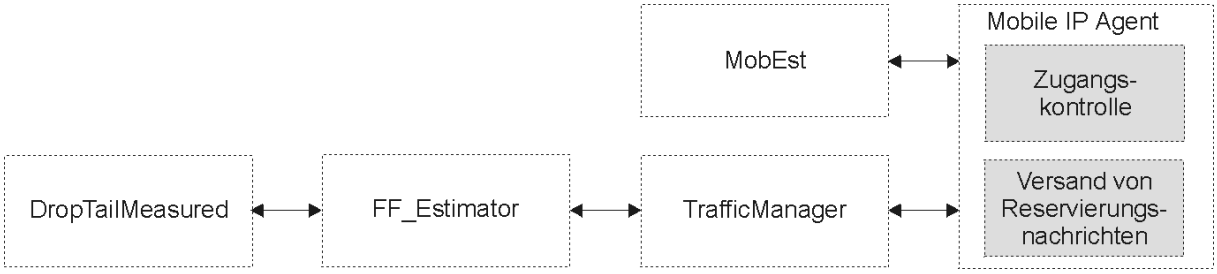

Abbildung 5.7: Integration von HoPVarB in ns-2

Bei **DropTailMeasured** handelt es sich um die Warteschlange des eingehenden Netzwerkinterface der Basisstation. Sie implementiert eine FIFO-Warteschlange, deren Ende bei einem Überlauf verworfen wird. Sie hat die Besonderheit, dass beim Einreihen eines Paketes ein Zähler um die Bytegröße des Paketes inkrementiert wird. Fragt man diesen Zähler periodisch ab, kann man so die einzelnen Messwerte hinsichtlich der Netzwerklast ermitteln. **FF_Estimator** implementiert das Modul zur Schätzung der Netzwerkauslastung, indem es die

Warteschlange jede Sekunde abfragt und daraus Auslastungsschätzungen errechnet. Der **TrafficManager** stammt aus [R01] und verwaltet einerseits Informationen über die lokale Basisstation: Gesamtbandbreite, belegte Bandbreite und Reservierungen der Nachbarstationen. Andererseits speichert er auch den Betrag der in den umliegenden Stationen verfügbaren Bandbreite ab. **MobEst** implementiert die Bewegungsvorhersage, sie wurde von [R01] übernommen. Wann immer der Mobile IP Agent einen Handover registriert, legt MobEst einen dementsprechenden Datensatz in der lokalen Datenbasis an. Im Gegensatz zu [R01] wird auch das Beenden einer Verbindung abgespeichert. Es wäre auch möglich, die Datenbasis beispielsweise auf dem Base Station Controller anzuordnen, der eine Gruppe von Basisstationen verwaltet. Wichtig ist, dass die Datenhaltung nicht zentral erfolgt, damit die Skalierbarkeit erhalten bleibt. Für HoPVarB wurde eine Anordnung direkt auf der Basisstation vorgesehen, um unnötige Kommunikationsvorgänge mit einer entfernten Datenbank zu vermeiden. Direkt in den **Mobile IP Agenten** wurden die Zugangskontrolle und der Versand der Reservierungsnachrichten eingefügt.

5.3.2 Integration der Zugangskontrolle in die Registrierung

Die in HoPVarB enthaltene Zugangskontrolle wurde in den Registrierungsvorgang von Mobile IP integriert. Dazu waren zwei Änderungen notwendig:

1. **Erweiterung des Paketkopfes** Der Mobile IP Paketkopf wurde um zwei Felder erweitert: Ein Feld für die beantragte mittlere Datenrate, und eines für die Adresse des alten Foreign Agent. Der Zweck des ersten Feldes ist unmittelbar klar. Es dient dazu, HoPVarB den Bandbreitenbedarf der neuen Verbindung mitzuteilen und wird somit von der Zugangskontrolle benötigt. Die Adresse des alten Foreign Agents dient wie in [R01] dazu, dass die Basisstation weiß, woher ein Mobile Node gekommen ist. Auf diese Weise lernt die Station ihre Nachbarn kennen.
2. **Erweiterung des Signalisierungsverfahrens** Um Zugangskontrolle und Handoverpriorisierung in Mobile IP integrieren zu können, musste der in Abschnitt 2.1.3 beschriebene Ablauf wie in [R01] erweitert werden. Die Entscheidung über die Zulassung der neuen Verbindung muss von demjenigen Agenten durchgeführt werden, in dessen Netzwerkbereich sich der Mobile Node befindet: Das können sowohl der Home Agent als auch der Foreign Agent sein. Der genaue Ablauf wird im Folgenden differenziert für die verschiedenen auftretenden Fälle beschrieben. Um die Zugangskontrolle in den Registrierungsvorgang integrieren zu können, benötigt der Foreign Agent allerdings Informationen, beispielsweise über den Auslastungszustand der Funkzelle. Möglich wird dies, indem man den Foreign Agent direkt auf der Basisstation anordnet. Problematisch an dieser Vorgehensweise ist, dass die zweite Betriebsart von Mobile IP so nicht mehr unterstützt werden kann, da sich der Foreign Agent auf dem mobilen Endgerät befindet und keine Information über den Zustand der

Basisstation hat. Ist diese Betriebsart erforderlich, müsste man ein Interface zwischen Basisstation und Foreign Agent definieren, über dass sie Informationen austauschen können.

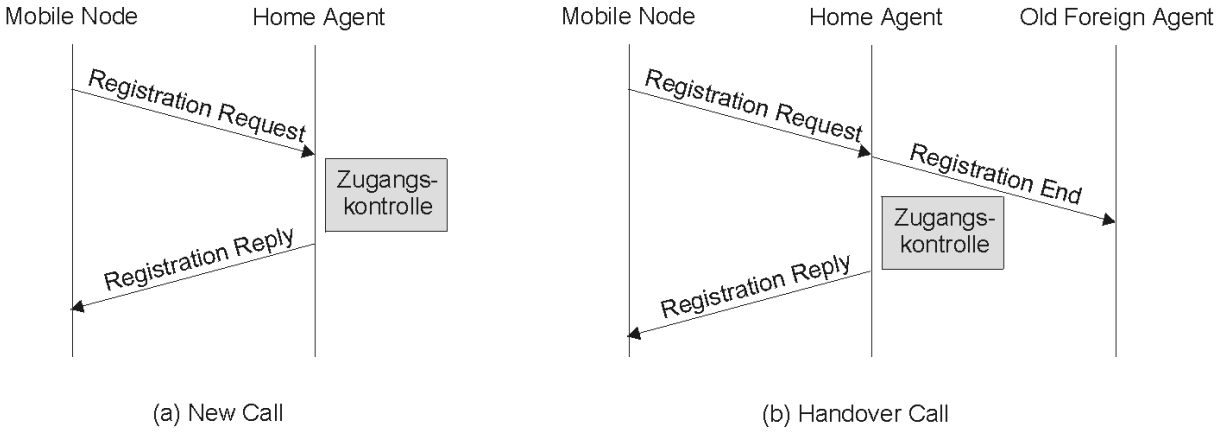

Abbildung 5.8: Registrierung eines Mobile Node beim Home Agent

Abbildung 5.8 zeigt den Registrierungsvorgang eines Mobile Nodes beim Home Agent in Form eines Weg-Zeit-Diagrammes. Da die Zugangskontrolle für New Calls und Handover Calls unterschiedlich erfolgt, sind auch die Registrierungsvorgänge verschieden. Im Falle eines New Calls (a) sendet der Mobile Host einen Registration Request mit der beantragten Bandbreite an den Home Agent. Dieser führt die Zugangskontrolle wie in Abschnitt 4.5.1 beschrieben durch und antwortet mit einem Registration Reply. Im Fall einer positiven Entscheidung enthält das Bandbreitenfeld den beantragten Wert, sonst wird es auf 0 gesetzt. Der Mobile Node weiß so, dass er abgewiesen wurde.

Im Fall eines Handover Calls (b) ist die Vorgehensweise geringfügig komplizierter. Der Mobile Node sendet einen Registration Request mit der gewünschten Bandbreite an den Home Agent. Anders als im ersten Fall sind zwei Dinge: Erstens läuft die Zugangskontrolle im Fall eines Handover anders ab (siehe auch Abschnitt 4.5.2). Zweitens muss noch eine Deregistrierung beim Old Foreign Agent erfolgen. Dabei handelt es sich um den Foreign Agent, bei dem der Mobile Node vorher registriert war. Auch im Handover-Fall erhält der Mobile Node einen Registration Reply mit der bewilligten Bandbreite.

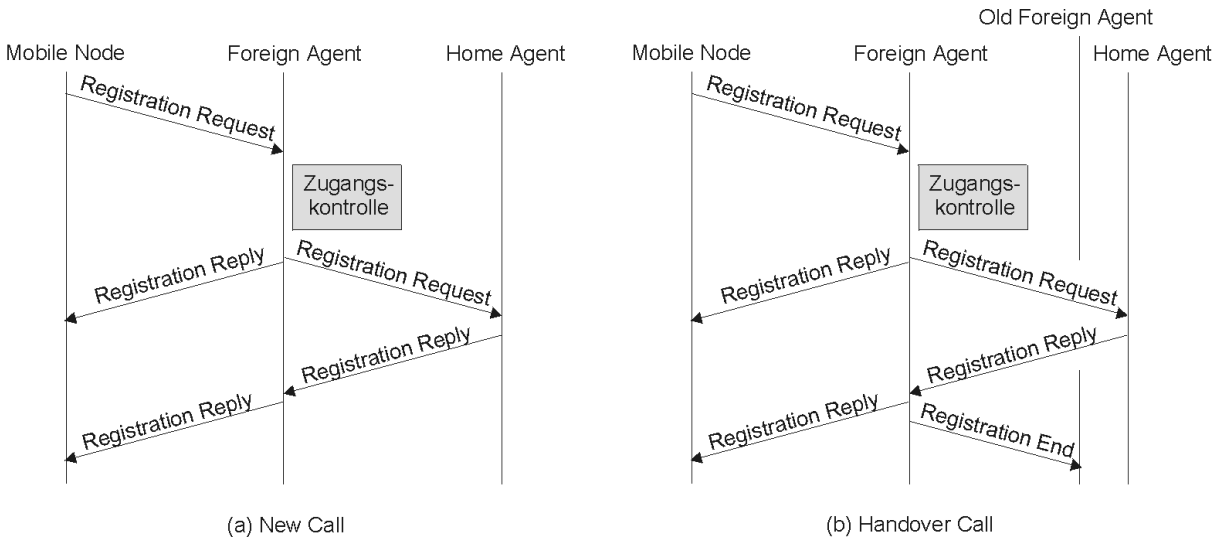

Abbildung 5.9: Registrierung eines Mobile Node beim Foreign Agent

Die Registrierung bei einem Foreign Agent ist umfangreicher (Abbildung 5.9). Der Mobile Node sendet im Fall eines New Calls (a) einen Registration Request mit der mittleren Bandbreite der Verbindung an den Foreign Agent. Nur hier wird die Zugangskontrolle durchgeführt. Fällt die Entscheidung negativ aus, wird lediglich ein Registration Reply mit der Bandbreitenangabe 0 an der Mobile Node geschickt, und die Signalisierung endet. Der Registration Request an den Home Agent wird also nur nach einer positiven Zulassungsentscheidung verschickt. In diesem Fall reserviert der Foreign Agent sofort die beantragten Ressourcen. Lehnt der Home Agent den Registration Request ab, werden diese nachträglich wieder freigegeben. Nach einem positiven Registration Reply des Home Agent wird dieser an den Mobile Node weitergeleitet.

Der Reservierungsablauf eines Handover Calls (b) unterscheidet sich abgesehen von der veränderten Zugangskontrolle nur dadurch von einem New Call, dass eine Registration End Nachricht an den alten Foreign Agent gesendet wird. Auf diese Weise wird die Registrierung bei dem Foreign Agent aufgehoben, bei dem der Mobile Node vorher gemeldet war.

5.4 Verkehrsmodelle

In den durchgeführten Simulationen kamen verschiedene Verkehrsquellen zum Einsatz. Solche Verkehrsgeneratoren haben die Aufgabe, Datenverkehr der Nutzer eines Netzwerkes zu simulieren. Um die Auswirkungen verschiedenen Nutzerverhaltens auf das Verfahren testen zu können, kamen Quellen mit unterschiedlichen Eigenschaften zum Einsatz, die nachfolgend mit ihren Parametern beschrieben werden.

5.4.1 Datenquellen mit konstanter Bitrate

Datenquellen mit konstanter Bitrate können eingesetzt werden, um beispielsweise unkomprimierten Audio- oder Videoverkehr zu simulieren.

Die in ns-2 implementierte Datenquelle mit konstanter Bitrate (constant bitrate, CBR) sendet beständig Pakete mit einer bestimmter Bitrate.

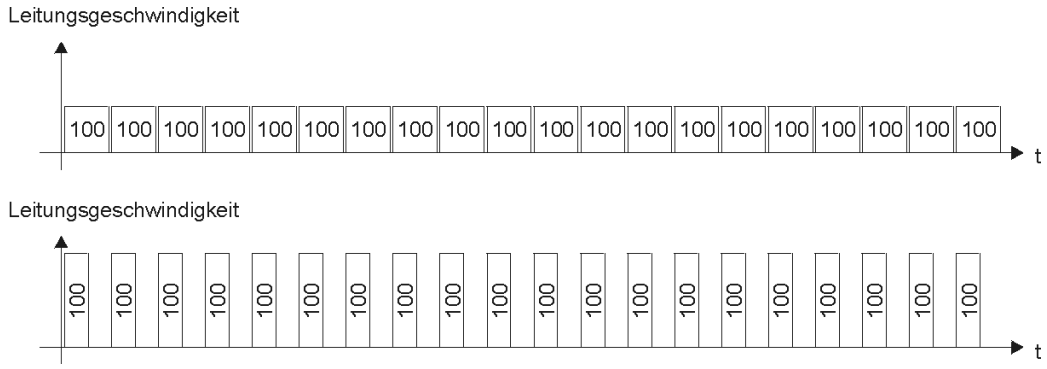

Abbildung 5.10: Verkehr einer Datenquelle mit konstanter Bitrate

Wie in Abbildung 5.10 dargestellt, haben alle Pakete eine voreingestellte Größe. Die Inter-Paketzeiten auf einer Leitung sind immer gleich, sie hängen von ihrer Geschwindigkeit ab. Ist die Bandbreite der Leitung höher als die Senderate der Quelle, entstehen Abstände zwischen den Paketen (unten); sind beide Werte gleich, folgen die Pakete direkt aufeinander (oben). Der erzeugte Datenverkehr hat somit keinerlei Schwankungen, mittlere und maximale Datenrate entsprechen sich. Überlagert man verschiedene solcher Datenquellen, entspricht die aggregierte Datenrate etwa der Summe der mittleren bzw. maximalen Datenraten der einzelnen Ströme. Dennoch können sich durch synchrones Senden mehrerer Quellen Schwankungen der maximalen Datenrate ergeben.

Um die Eigenschaften einer solchen Datenquelle zu bestimmen, sind die Paketgröße und die Datenrate festzulegen.

5.4.2 Datenquellen mit variabler Bitrate

Datenquellen mit variabler Bitrate (Variable Bitrate, VBR) eignen sich beispielsweise für die Simulation von komprimierten Audio- oder Videodatenströmen. Sie haben die Eigenschaft, dass ihre Datenrate über die Zeit variiert. Nach oben hin ist sie durch die maximale Datenrate (Peakrate) begrenzt. Im Durchschnitt wird mit einer mittleren Datenrate gesendet, die unterhalb der Peakrate liegt.

VBR-Datenquellen werden meist durch sogenannte On-Off-Datenquellen approximiert. Dabei schaltet die Quelle zwischen zwei Betriebsarten um: Während sie sich im On-Modus wie eine CBR-Datenquelle verhält, sendet sie im Off-Modus keinerlei Daten (Abbildung 5.11).

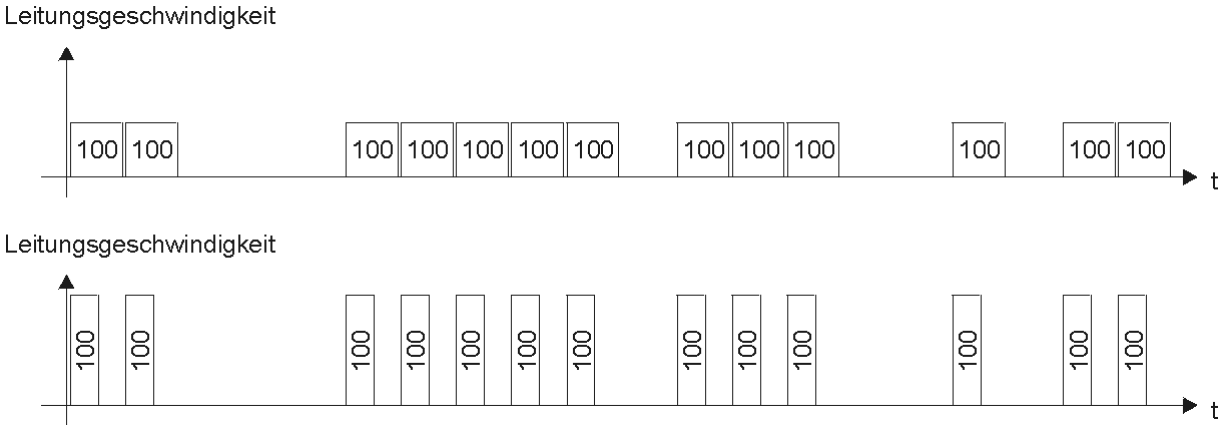

Abbildung 5.11: Verkehr einer On-Off-Datenquelle

Dabei wird das Verhalten der Datenquelle maßgeblich durch die Länge der einzelnen On- und Off-Intervalle beeinflusst. Sie werden einer Zufallsvariablen entnommen und variieren um einen festgelegten Mittelwert (den sogenannten Erwartungswert). Durch das Verhältnis von mittlerer On- und Off-Zeit (t_{on}/t_{off}) wird festgelegt, wie stark die Datenrate variiert. Daraus ergibt sich, wie groß der Unterschied zwischen der mittleren Datenrate r_{avg} und der Peakrate r_{max} im On-Modus ist. Die mittlere Datenrate ergibt sich aus

$$r_{avg} = \frac{t_{on}}{t_{on} + t_{off}} r_{\max}.$$

Überlagert man verschiedene On-Off-Datenquellen, so hat der aggregierte Datenstrom wiederum eine variable Bitrate. Seine mittlere Bitrate entspricht der Summe der einzelnen mittleren Raten. Das liegt daran, dass sich Sende- und Ruhephasen der einzelnen Quellen ausgleichen. Theoretisch ist die maximale Datenrate des aggregierten Stromes gleich der Summe der einzelnen Peakrates: Es wäre möglich, dass zu irgendeinem Zeitpunkt alle Quellen gleichzeitig mit voller Datenrate senden. In der Praxis ist allerdings davon auszugehen, dass die aggregierte Peakrate nur wenig über der mittleren aggregierten Bitrate liegt, da sich die Spitzenwerte herausmitteln.

Über die beiden CBR-Parameter Senderate und Paketgröße hinaus sind bei Quellen mit variabler Bitrate zusätzlich der Erwartungswert für On- und Off-Periode festzulegen. Darüber hinaus hat die Art der Zufallsvariablen Einfluss: Je nach deren Verteilung ändern sich die Eigenschaften der einzelnen Ströme sowie des Aggregates.

ns-2 implementiert unter anderem VBR-Quellen, die auf Exponential- und Paretoverteilungen basieren. Der exponentialverteilten Quelle (Exponential On Off, EXPOO) wurden die Veränderungen aus [FL00] hinzugefügt, um ein klar definiertes Sendeverhalten zu erreichen und so die Vergleichbarkeit mit anderen EXPOO-Implementierungen sicherzustellen.

Die paretoverteilten Quellen (Pareto On Off, POO) benötigen einen weiteren Parameter β, der die Form der Verteilung beeinflusst. Er wurde in Anlehnung an [JSD97] auf 1.2 bzw. 1.9

gesetzt. Ist $\beta \leq 2$, hat die Verteilung eine unendliche Varianz, für $\beta \leq 1$ hat sie einen unendlichen Mittelwert. Überlagert man mehrere paretoverteilte Quellen (Pareto On Off, POO), so ist im Gegensatz zur Exponentialverteilung der entstehende Verkehr selbstähnlich ([PF95]). Das liegt daran, dass Paretoverteilungen *heavy tailed* sind. Das bedeutet formal, dass die *conditional mean exceedance* (CME_x) eine stetig steigende Funktion von x ist. Die Bedeutung dessen lässt sich in einem Beispiel veranschaulichen: Wäre die Wahrscheinlichkeitsverteilung von Wartezeiten in einem Arztwartezimmer heavy tailed, würde das bedeuteten, dass die zu erwartende verbleibende Wartezeit größer würde, je länger man bereits gewartet hat. Ist die Verteilung nicht heavy tailed, ist die verbleibende Wartezeit von der bisherigen Wartezeit unabhängig.

5.5 Interpretation der Simulationsergebnisse

In diesem Abschnitt sollen die durchgeführten Simulationen und die dabei ermittelten Ergebnisse beschrieben und interpretiert werden.

Um zu einer Bewertung der Simulationsergebnisse zu kommen, wurde die in [HR99] vorgestellte Kostenfunktion *Grade of Service* (GoS) verwendet. Sie bezieht New Call Blocks (NCB) wie Handover Drops (HOD) in Relation zu Verbindungswünschen (N) bzw. erfolgreich angenommenen Verbindungen (New Calls, NC) in die Betrachtung ein. Da Handover Drops als schwerwiegendere Dienstgütemängel angesehen werden als New Call Blocks, werden sie mit dem Faktor 5 gewichtet. Somit berechnet sich der Grade of Service wie folgt:

$$GoS = \frac{NCB}{N} + 5\frac{HOD}{NC}.$$

Mit steigender Netzlast wird GoS ansteigen, da prozentual weniger Verbindungen zugelassen werden können. Bei starker Netzüberlastung wird sich der erste Summand von unten der 1 nähern, es gilt

$$\frac{NCB}{N} = \frac{N - NC}{N} \xrightarrow{N \to \infty} 1,$$

da die Anzahl der zugelassenen New Calls (NC) wegen der begrenzten Netzkapazität etwa konstant bleibt. Gelingt es HoPVarB, die Anzahl der Handover Drops zu begrenzen, müsste somit die Funktion des GoS über der Last qualitativ in etwa den in Abbildung 5.12 dargestellten Verlauf haben. Eine Last von 100% bedeutet dabei, dass die Summe der mittleren Bitraten der *N* Verbindungen genau der Bandbreitenkapazität der Basisstationen entspricht.

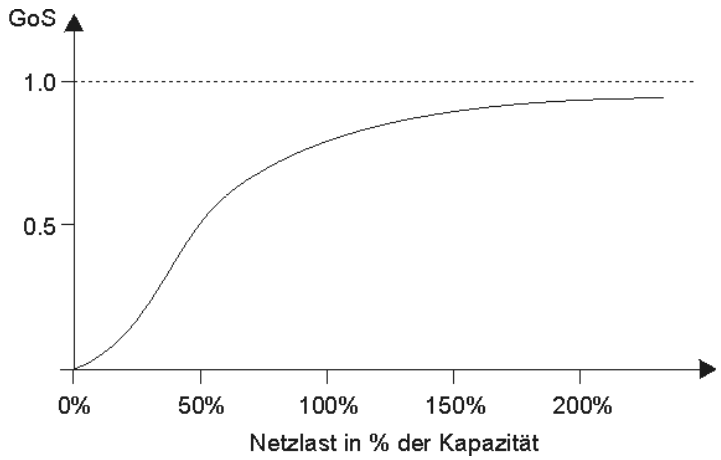

Abbildung 5.12: Erwarteter Verlauf der GoS-Kurve

Dabei gilt grundsätzlich, dass das Verfahren umso positiver zu bewerten ist, je flacher der Verlauf der GoS-Kurve ist. Im direkten Vergleich zweier Verfahren ist dasjenige vorzuziehen, das bei gleichen Bedingungen die niedrigeren GoS-Werte erzielt. Ein Verfahren zur Handoverpriorisierung muss als gescheitert angesehen werden, wenn die GoS-Kurve bei steigender Last nicht durch 1.0 nach oben beschränkt ist, sondern darüber hinaus wächst.

5.5.1 Datenströme mit variabler Bitrate

Zunächst soll die prinzipielle Leistungsfähigkeit von HoPVarB untersucht werden. Es wurden die GoS-Werte für unterschiedliche Netzlasten ermittelt. Auf diese Weise wurde der Grade of Service als Funktion der Netzlast dargestellt.

Dabei kamen Datenströme mit variabler Bitrate zum Einsatz, die aus einer exponentialverteilten Quelle kamen. Die maximale Datenrate betrug mit 8 KBit/s das Doppelte der mittleren Datenrate. Der Parameter Reservierungshöhe war auf 1000 Promille eingestellt, die simulierte Zeit betrug 33 Minuten.

Es wurden jeweils zwei Simulationsläufe durchgeführt und die Ergebnisse gemittelt, um simulationsbedingten Schwankungen entgegenzuwirken. Die Untersuchung der Schwankungen der beiden (hinterher gemittelten) Ergebnisse ergab, dass die ermittelten GoS-Werte sich um etwa 3-4% unterscheiden.

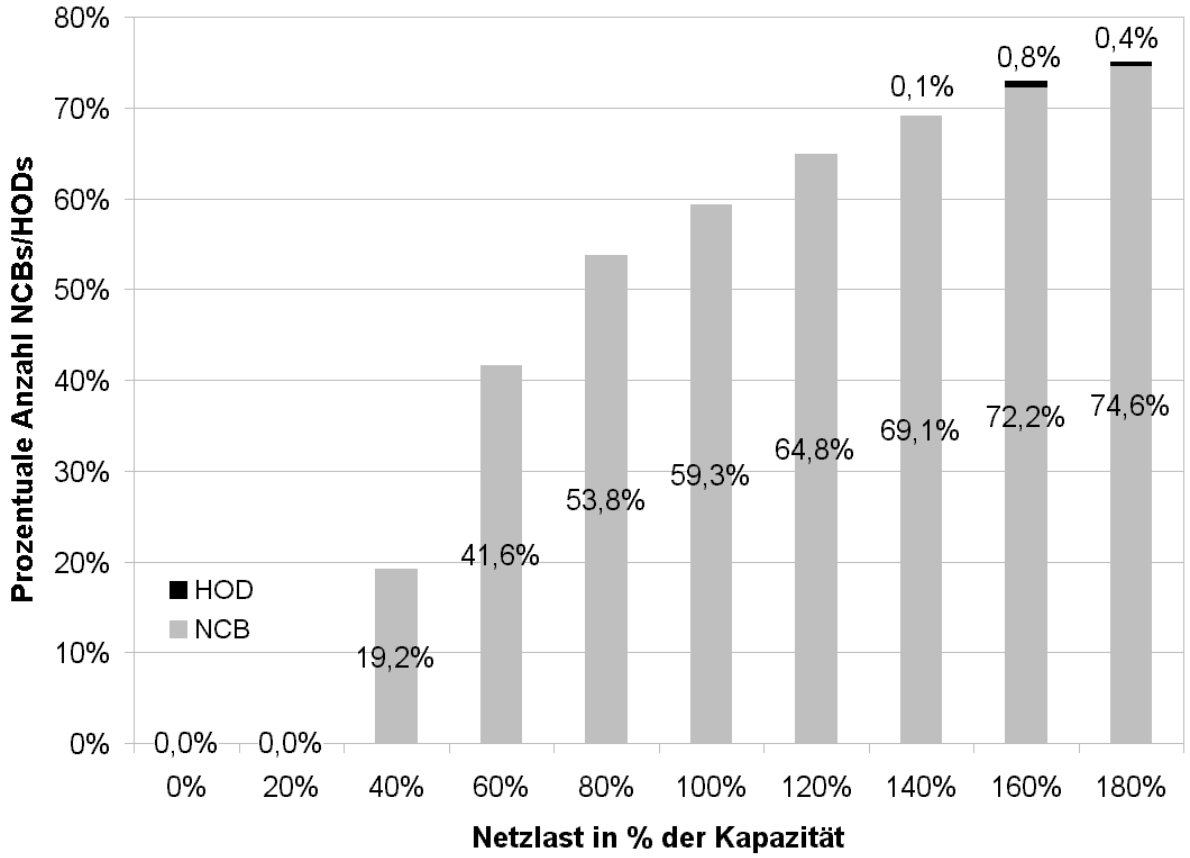

Abbildung 5.13: NCBs und HODs (prozentual) bei Datenströmen mit variabler Bitrate

Die Ergebnisse sind in Abbildung 5.13 dargestellt. Dabei ist die Anzahl der New Call Blocks in Relation zur Anzahl der Verbindungswünsche als grauer Balken dargestellt, während die Anzahl der Handover Drops in Relation zur Anzahl der begonnenen Gespräche schwarz dargestellt ist.

Es ist zu erkennen, dass der Anteil der abgewiesenen Verbindungen erst schnell steigt, um dann mit weiter wachsender Last langsamer zuzunehmen. Er nähert sich der 100%-Marke asymptotisch. Das liegt daran, dass aufgrund der begrenzten Netzkapazität die absolute Anzahl der zugelassenen Verbindungen bei zunehmender Last etwa gleich bleibt, die Anzahl der Verbindungswünsche jedoch steigt (siehe auch Abbildung 5.14 (a)).

Bemerkenswert ist in diesem Zusammenhang, dass es erst ab 140% Last überhaupt zu Handover Drops kommt. Es kam bei höchstens 0,8% der zugelassenen Gespräche zu Abbrüchen wegen eines Zellwechsels. Diese Werte können als konservativ eingestuft werden. Daher ist zu überlegen, inwieweit durch eine Reduzierung des Steuerparameters Reservierungshöhe die Netzauslastung gesteigert werden kann, ohne dass die Rate der Handover Drops nennenswert steigt.

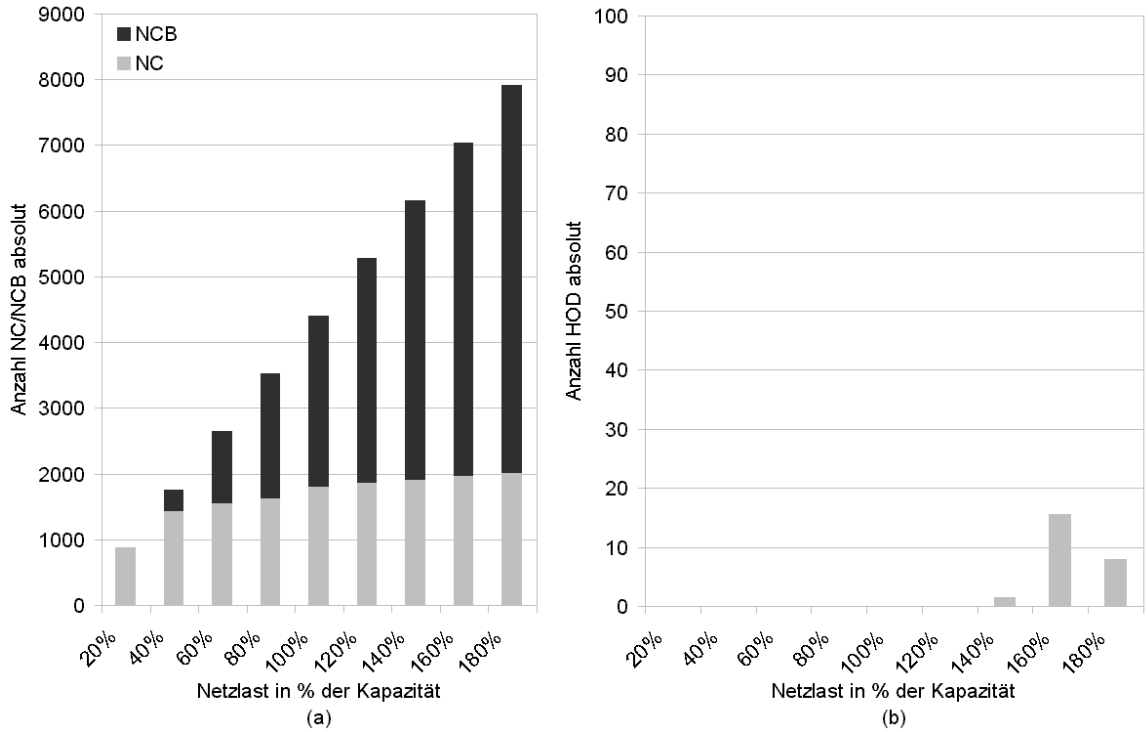

Abbildung 5.14: NCs, NCBs und HODs (absolut) bei Datenströmen mit variabler Bitrate

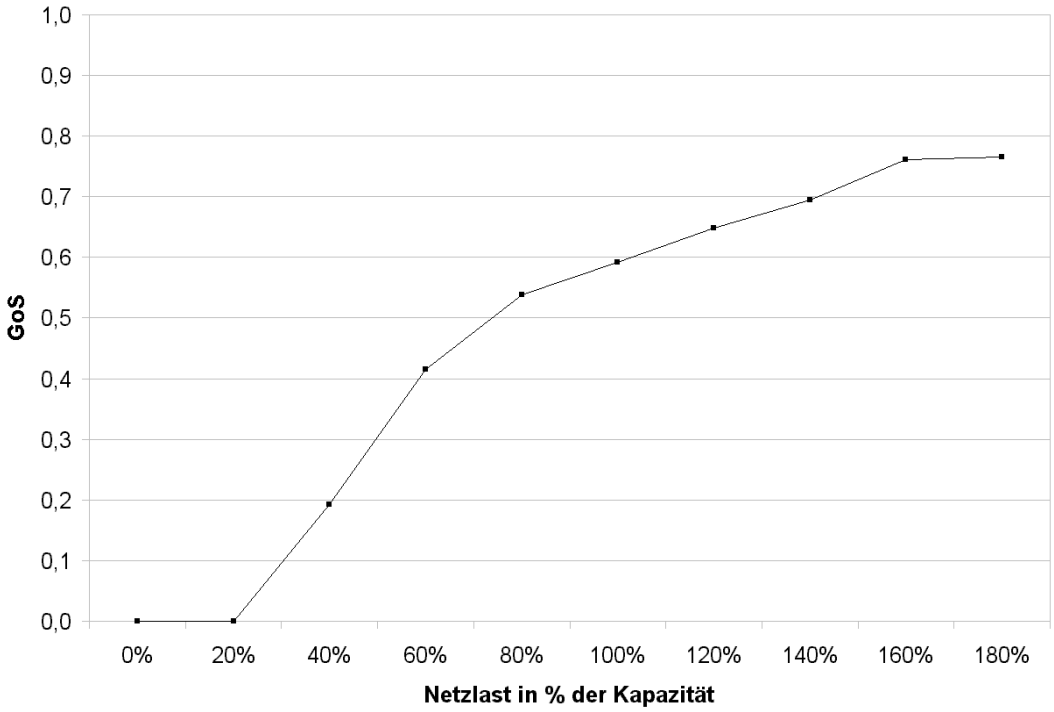

Abbildung 5.15: GoS-Kurve bei Datenströmen mit variabler Bitrate

In Abbildung 5.15 ist die sich aus den prozentualen Werten in Abbildung 5.13 ergebende GoS-Kurve dargestellt. Es ist zu erkennen, dass sie den erwarteten Verlauf nimmt. Bei geringen Auslastungen sind die GoS-Werte niedrig, da ein großer Anteil der New Calls zugelassen werden kann und es praktisch nicht zu Handover Drops kommt. Die Kurve steigt

jedoch im weiteren Verlauf steil an, da der Anteil der zugelassenen New Calls zu Gunsten von Reservierungen für Handovers abnimmt. Ab einer Last von 60% verläuft die Kurve dann zunehmend flacher. Wie oben bereits beschrieben, liegt das daran, dass der Anteil der New Call Blocks sich langsam 100% nähert, während Handover Drops effektiv begrenzt werden.

Es ist interessant, dass der Wendepunkt der Kurve in Abbildung 5.15 etwa bei 60% Last erkennbar ist. Jeder zugelassene Datenstrom belegt nicht nur für die Übertragung Ressourcen in Höhe seiner mittleren Datenrate, sondern nahezu den gleichen Wert an Reservierungen in den umliegenden Zellen. Dabei haben die Reservierungen nicht exakt den gleichen Wert wie die belegten Ressourcen, da im Mittel 17% der Verbindungen einer Zelle diese nicht als Handover verlassen, sondern in dieser Zelle enden. Deshalb werden für den von ihnen genutzten Ressourcenanteil keine Reservierungen in den umliegenden Zellen vorgenommen. Daher belegt jede Verbindung insgesamt etwa das 1,83-fache ihrer mittleren Datenrate an Ressourcen im Netz. Somit liegt die Auslastungsgrenze des Netzes aufgrund der Reservierungen bei 54,6%. Das bedeutet, dass das Netz bei einer Reservierungshöhe H=1000 überhaupt nur zu etwas mehr als der Hälfte ausgelastet werden kann.

5.5.2 Einfluss des Steuerparameters Reservierungshöhe

Wie bereits angesprochen, kann die Reservierungshöhe über den Steuerparameter *H* manuell in Promilleschritten angepasst werden. Auf diese Weise ist es möglich, das Ausmaß der Priorisierung von Handovers gegenüber New Calls zu beeinflussen.

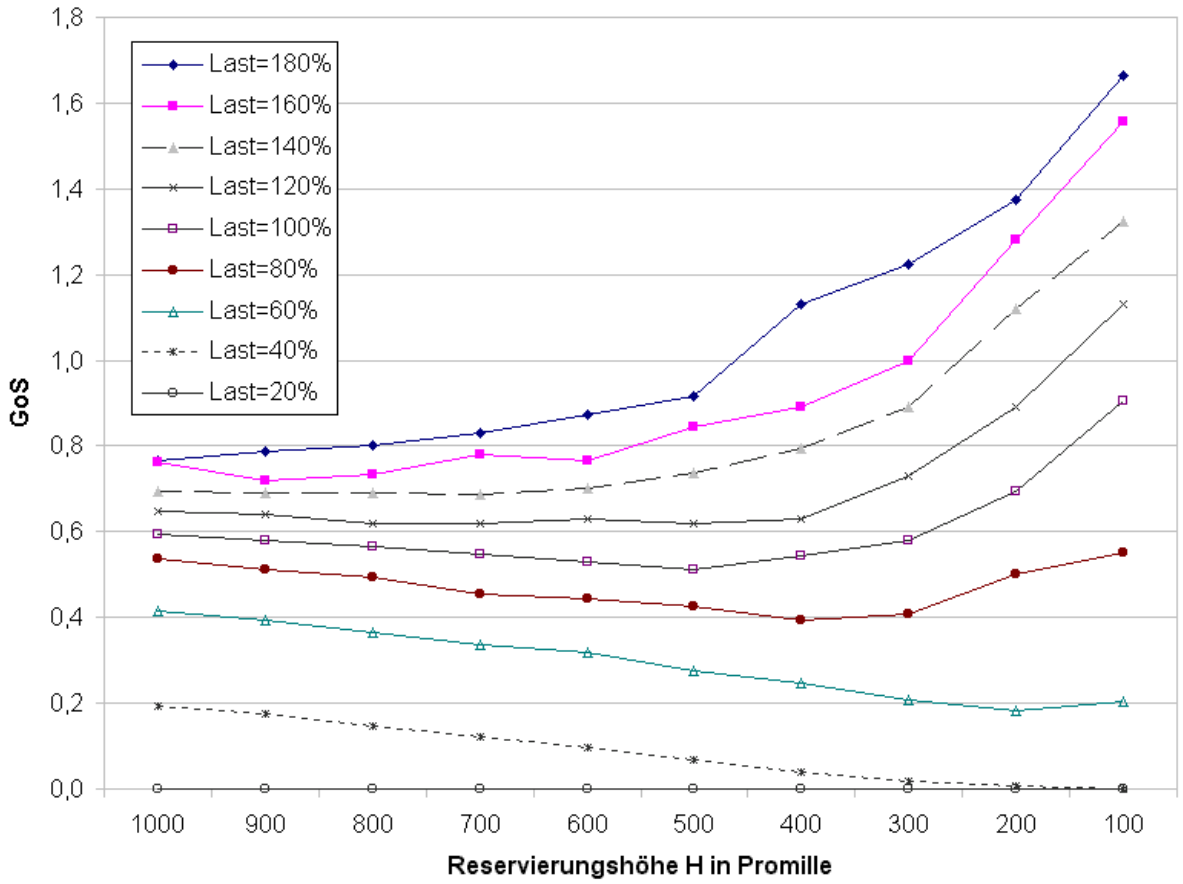

Abbildung 5.16: Grade of Service bei unterschiedlichen Reservierungshöhen

Abbildung 5.16 zeigt die Auswirkungen unterschiedlicher Reservierungshöhen auf die GoS-Werte in verschiedenen Lastsituationen. Abgesehen von der Variation des Steuerparameters entspricht das Szenario dem aus Abschnitt 5.5.1; somit entspricht der erste Wert einer jeden Kurve (100%) den in Abbildung 5.15 gezeigten Werten.

Es ist zu erkennen, dass sich die Variation der Reservierungshöhe je nach Netzlast unterschiedlich auswirkt: Bei hoher Last werden die GoS-Werte mit abnehmender Reservierungshöhe deutlich größer (d.h. schlechter), die Grenze bei 1.0 wird deutlich überschritten. Eine mittlere Last führt dazu, dass sich bei einer bestimmten Reservierungshöhe ein Optimum bildet. Bei geringer Last hingegen ist der GoS-Wert umso niedriger (und somit besser), desto geringer die Reservierungshöhe ist.

In Abbildung 5.17 ist eine Erklärung für diese Tatsache zu erkennen. Sie zeigt die GoS-Kurve bei einer Last von 180% (a), 100% (b) sowie 40% (c). Dabei wurden die beiden Summanden der GoS-Funktion jeweils getrennt dargestellt. Der erste Teil, der auf New Call Blocks zurückgeht, ist hellgrau dargestellt, während der durch Handover Drops beigesteuerte Anteil dunkelgrau gefärbt ist.

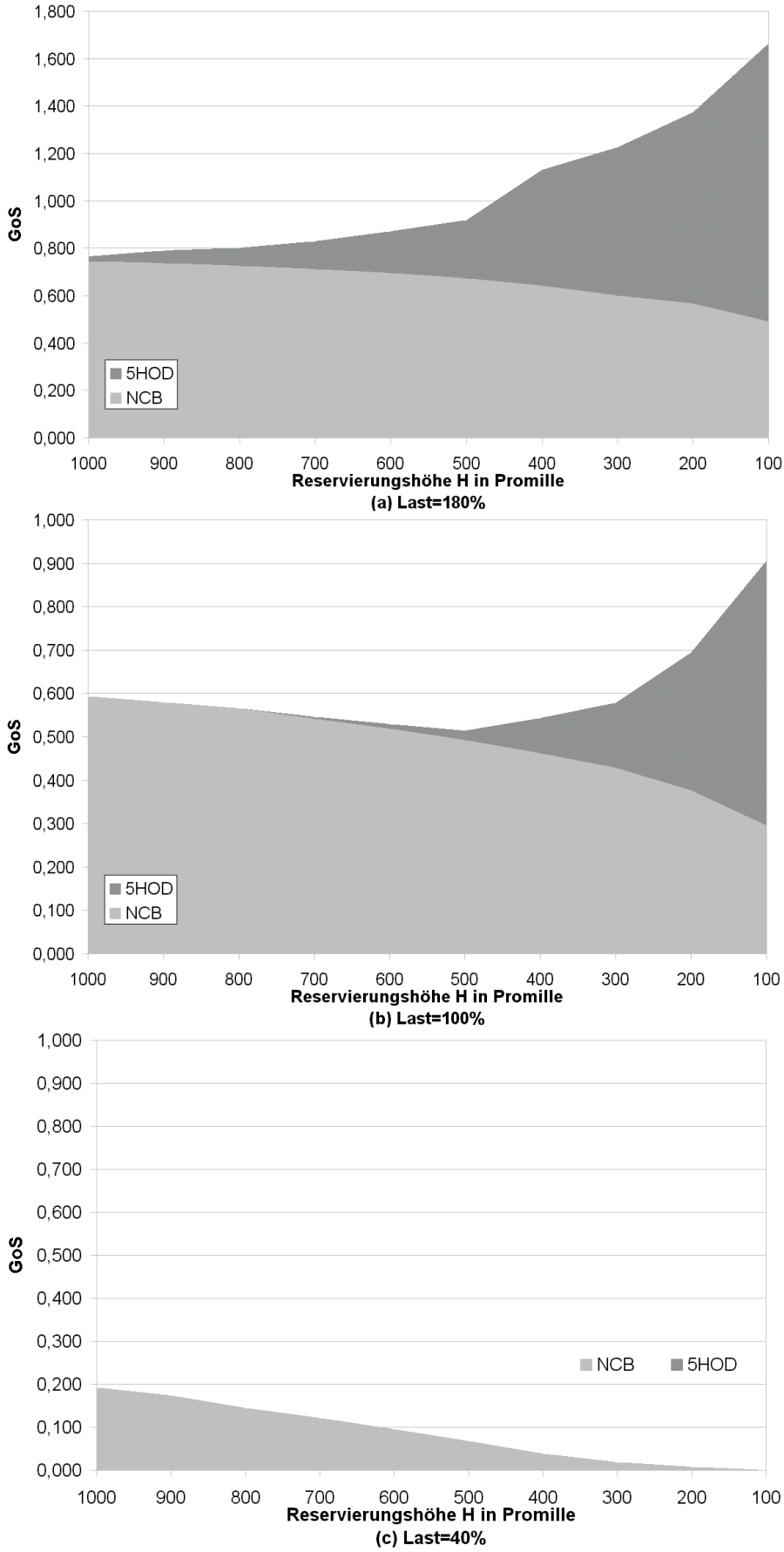

Abbildung 5.17: Grade of Service differenziert nach NCB und HOD Rate

Es ist deutlich zu erkennen, wie bei einer Last von 180% der erste Teil mit zurückgehenden Reservierungen abnimmt. Dies ist nicht überraschend, da weniger Bandbreite für Handovers reserviert wird, und somit mehr Ressourcen für die Zulassung neuer Verbindungen zur Verfügung stehen. Folglich müssen weniger New Calls abgewiesen werden. Dies geht jedoch im Fall (a) deutlich auf Kosten der Handover Drops: durch die reduzierten Reservierungen steht zunehmend nicht genügend Bandbreite für Handovers zur Verfügung, die absolute Anzahl der Handover Drops steigt. Das führt dazu, dass der zweite Teil der GoS-Formel zunimmt. Da dieser durch das Gewicht deutlich betont wird, verschlechtert sich der GoS-Wert insgesamt bei abnehmenden Reservierungen.

Bei 100% Last ist die Situation ähnlich. Während die New Call Blocks und somit der erste Teil der GoS-Formel mit rückläufigen Reservierungen abnehmen, nimmt die Anzahl der Handover Drops und somit der zweite Teil der Formel ab einer gewissen Reduzierung zu. Ein wesentlicher Unterschied zu (a) ist jedoch, dass die Handover Drops erst ab einer Reservierungshöhe von 40% nennenswert steigen, vorher wirkt sich die Verringerung der Reservierungen kaum aus. Das führt dazu, dass die GoS-Kurve insgesamt zunächst fällt und dann wieder steigt: Es bildet sich ein Optimum bei einer Reservierungshöhe von 50%.

Anders ist die Situation bei einer Last von 40% der Netzkapazität. Da ein Teil der Ressourcen durch die geringe Last verfügbar ist, ist es nicht erforderlich, Reservierungen vorzunehmen: Die Ressourcen sind ohnehin nicht belegt. Daher kommt es auch bei deutlich verminderten Reservierungen nicht zu Handover Drops. Dennoch kann die Anzahl der zugelassenen Verbindungen gleichzeitig gesteigert werden, der GoS-Wert sinkt. Das liegt daran, dass die Gesamthöhe der Reservierungen der umliegenden Basisstationen in den verschiedenen Zellen unterschiedlich ist.

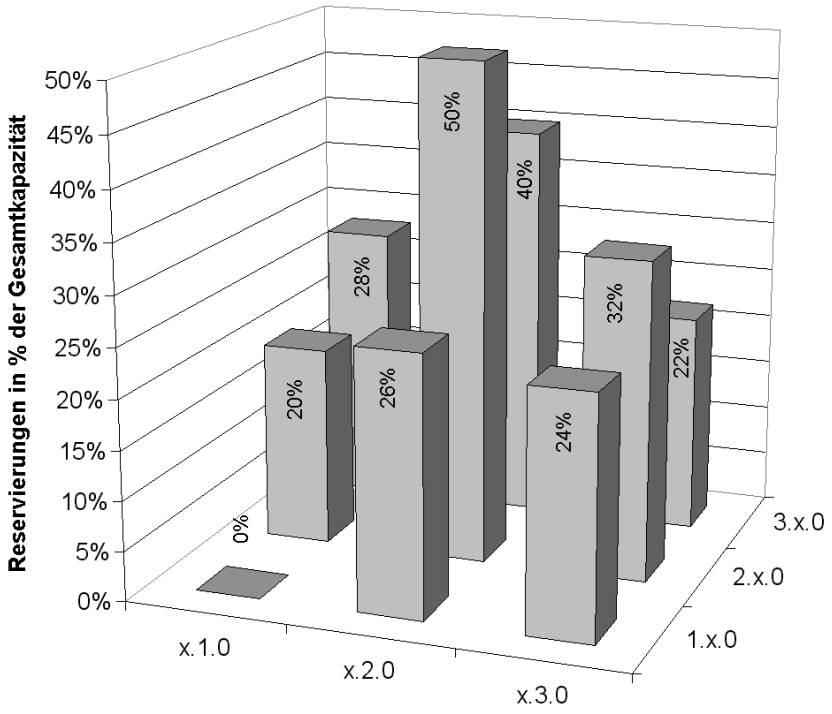

Abbildung 5.18: Reservierungshöhen in den verschiedenen Basisstationen

Abbildung 5.18 zeigt den Betrag der Reservierungen in Prozent der Gesamtkapazität in den neun Basisstationen zu Beginn der zweiten Simulationshälfte (Sekunde 1228). Die Last betrug 40%, der Steuerparameter Reservierungshöhe ist auf 1000 Promille gesetzt. Wie bereits in Abschnitt 5.1.3 beschrieben, ist der Betrag in Zelle 2.2.0 am größten, da hier für Mobilfunkteilnehmer aus allen vier Nachbarzellen Reservierungen vorgenommen werden müssen. In einer solchen Zelle können bei verringerter Reservierungshöhe *H* mehr New Calls zugelassen werden; der betreffende Anteil des GoS-Wertes verringert sich. Die Zelle 1.1.0 ist nicht mit dem Straßennetz verbunden. Somit kommt niemals ein Teilnehmer zu der Basisstation, folglich müssen niemals Reservierungen vorgenommen werden.

Insgesamt wird deutlich, dass die optimale Reservierungshöhe von der zu erwartenden Netzlast abhängt. Betrachtet man Abbildung 5.16, wird klar: bei hoher Last ist eine Reservierungshöhe von 1000 Promille am günstigsten, bei geringer Last eine Reservierungshöhe von beispielsweise 100 Promille. Es ist aber davon auszugehen, dass beide Lastsituationen selten sind. Geht man von Lastsituationen um 100% aus, erzielt eine Reservierungshöhe von 700 – 800 Promille die günstigsten Werte: bei keiner der benachbarten Kurven steigt bis zu diesem H die Kurve gravierend an, es lassen sich in den unteren Lastbereichen auf diese Weise gute Ergebnisse erzielen.

Eine weitere Möglichkeit, sich für ein Reservierungsniveau zu entscheiden, stellt die Festlegung einer Obergrenze für die prozentuale Anzahl der Handover Drops dar.

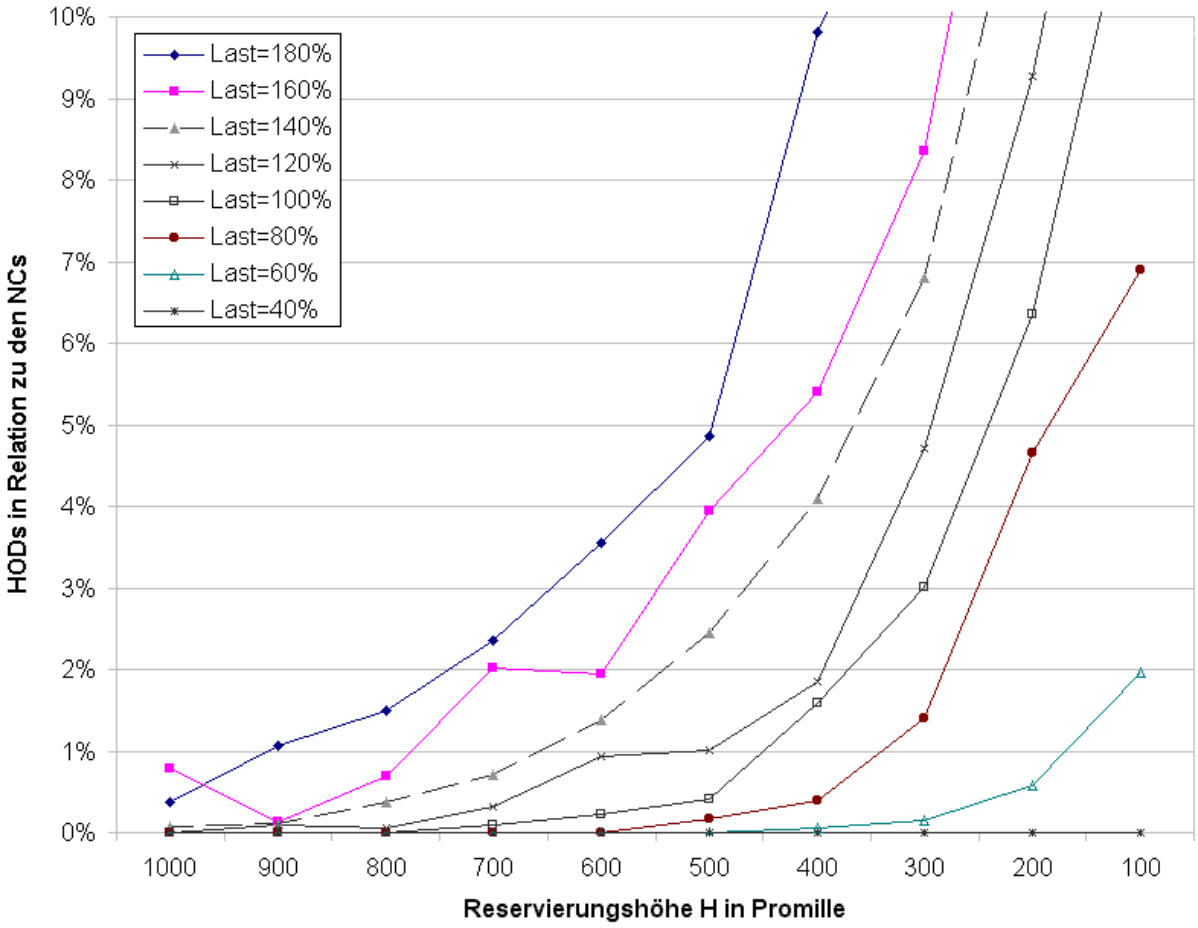

Abbildung 5.19: Prozentuale Anzahl von Handover Drops

In Abbildung 5.19 ist die Anzahl der Handover Drops in Relation zu der Anzahl der New Calls in Abhängigkeit von Netzlast und Reservierungshöhe abgebildet. Die Kurven entsprechen somit dem zweiten Teil der GoS-Formel, allerdings ohne das Gewicht. Dargestellt ist somit

$$\frac{HOD}{NC}.$$

Dies stellt ein Maß dafür dar, in wie viel Prozent der begonnenen Gespräche es zu einem Verbindungsabbruch kommt, der durch einen Zellenwechsel bedingt ist.

Je nach dem, wie viel Prozent Handover Drops man zu akzeptieren bereit ist, kann man nun anhand von Abbildung 5.19 die passende Reservierungshöhe festlegen. Wird ein Prozent als akzeptabel angesehen, so muss die Reservierungshöhe auf 900 Promille eingestellt werden: Etwa bei H=90% schneidet die Last=180%-Kurve die waagerechte 1%-Linie. Ist man bereit, zwei Prozent Handover Drops hinzunehmen, reicht eine Reservierungshöhe von 750 Promille aus. Ignoriert man eine Last von 180% als Sonderfall und sieht 160% Netzlast als Obergrenze an, reduzieren sich die anzustrebenden Reservierungshöhen auf knapp 800 bzw. 600 Promille.

5.5.3 Vergleich mit dem Peakrate-Ansatz

Eine wesentliche Neuerung von HoPVarB im Gegensatz zu der Handoverpriorisierung in [R01] ist, dass der Peakrate-Ansatz durch ein Modul zur Schätzung der Netzwerkauslastung ersetzt wurde. Es ist zu erwarten, dass bei Verkehr mit variabler Bitrate die Netzlauslastung deutlich steigt, und somit der GoS-Wert von HoPVarB deutlich unter dem des Peakrate-Ansatzes liegt. Das liegt daran, dass in letzterem bei Ressourcenbelegung und Reservierungen immer die deutlich höhere maximale Datenrate angesetzt wird. Bei Verkehr mit konstanter Bitrate hingegen ist zu erwarten, dass HoPVarB bestenfalls genauso gut abschneidet wie das Peakrate-Verfahren. Da sich in diesem Fall Peakrate und Average Rate nicht unterscheiden, legen hier beide Verfahren in etwa die gleiche Bitrate zugrunde. Es ist sogar möglich, dass HoPVarB beispielsweise durch Messfehler geringfügig schlechter abschneidet.

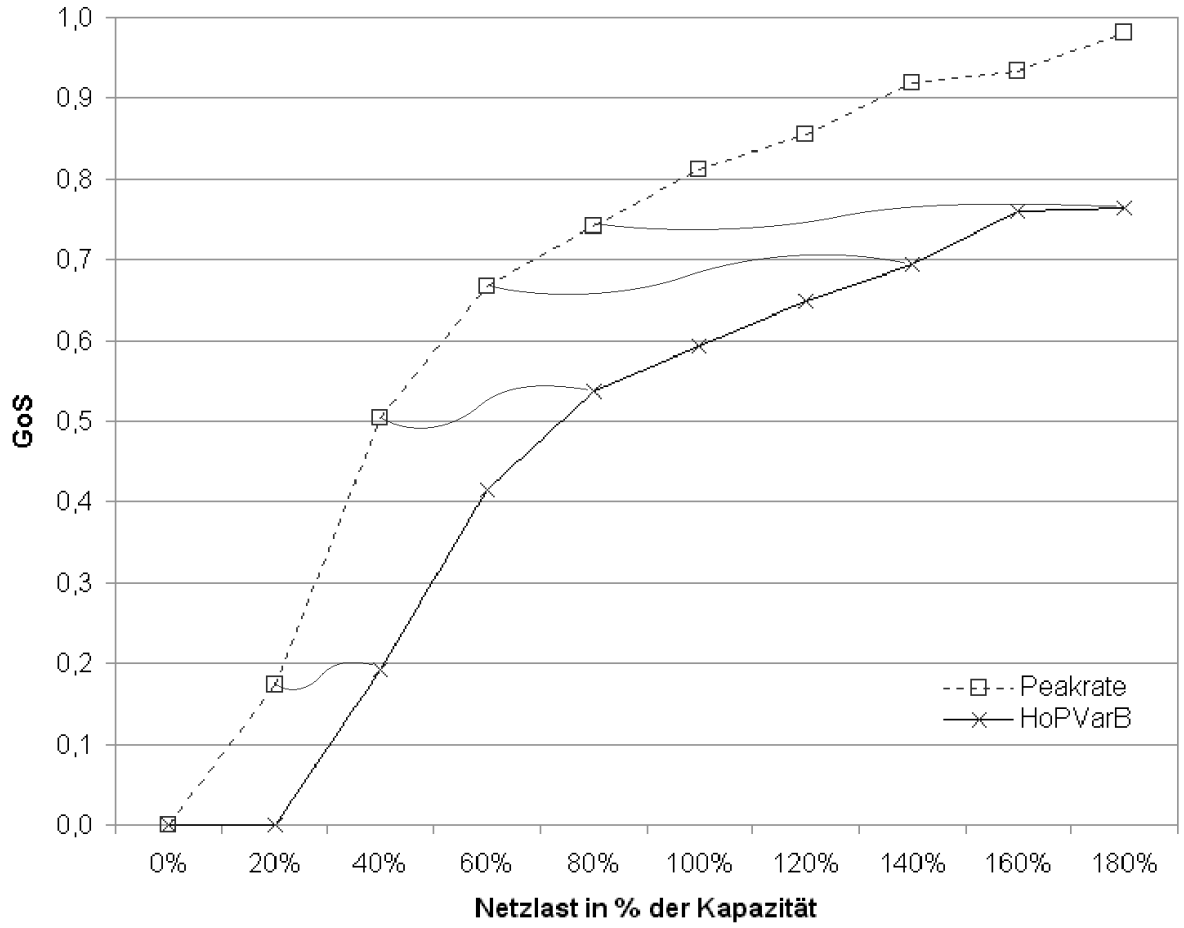

Abbildung 5.20: GoS von Peakrate-Ansatz und HoPVarB bei variabler Bitrate

Abbildung 5.20 zeigt die GoS-Kurven des Peakrate-Verfahrens und von HoPVarB bei Datenströmen mit variabler Datenrate im Vergleich. Es liegt das gleiche Szenario wie in Abschnitt 5.5.1 zugrunde, die maximale Datenrate lag mit 8 KBit/s doppelt so hoch wie die mittlere Datenrate. Erwartungsgemäß liegt die Kurve von HoPVarB deutlich unter der des Peakrate-Verfahrens. Das liegt daran, dass letzteres etwa den doppelten Ressourcenverbrauch

hat, da die höhere Peakrate zugrunde gelegt wird. Es ist zu erkennen, dass das Peakrate-Verfahren bei einer bestimmten Last etwa den GoS-Wert erreicht (z.B. Last=40%, GoS=0,494), den HoPVarB bei der doppelten Last erzielt (z.B. Last=80%, GoS=0,529). Die betreffenden Knoten sind in dem Diagramm mittels einer geschwungenen grauen Linie verbunden.

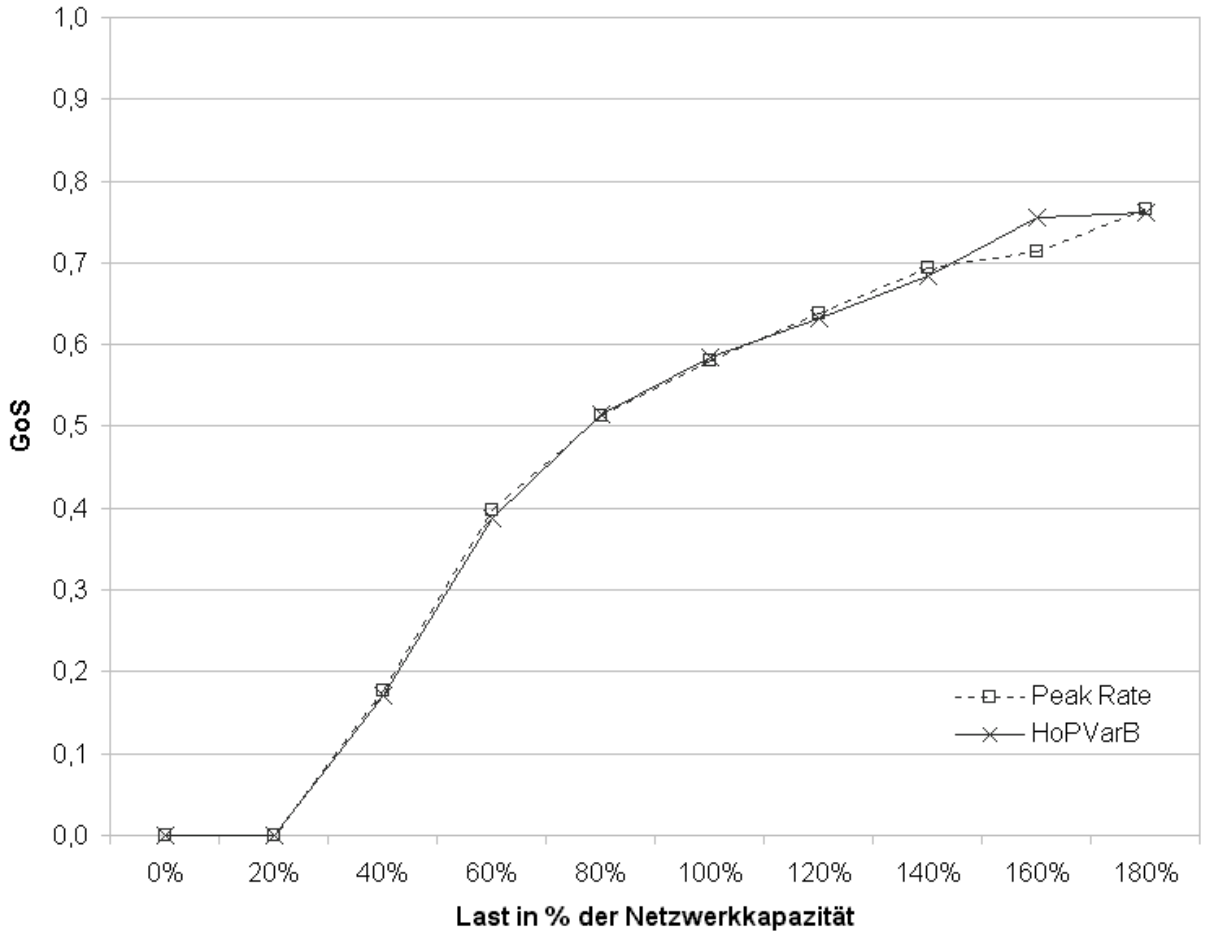

Abbildung 5.21: GoS von Peakrate-Ansatz und HoPVarB bei konstanter Bitrate

Anders ist die Situation, wenn Datenströme mit konstanter Bitrate simuliert werden. In Abbildung 5.21 ist sichtbar, dass hier beide Verfahren in etwa gleichauf liegen. In Abweichung zum vorhergehenden Szenario kamen Datenquellen mit einer konstanten Datenrate von 4 KBit/s zum Einsatz. Da nun kein Unterschied mehr zwischen maximaler und mittlerer Datenrate vorliegt, veranschlagen beide Verfahren in etwa den gleichen Ressourcenbedarf. Folglich sind die Unterschiede in der Leistungsfähigkeit gering.

Es ist bemerkenswert, dass durch die Messungen kein nennenswerter „Reibungsverlust“ durch Messfehler oder ähnliches zustande kommt. Einzig bei einer Last von 160% gibt es eine Abweichung von 0,043. Die Tatsache, dass die Messwerte bis zu 10% über den tatsächlichen Werten liegen können, wirkt sich auf die Gesamtleistungsfähigkeit des Verfahrens nicht aus. Das liegt zum einen daran, dass die Messungen nur einer von mehreren

Einflussfaktoren sind, zum anderen aber auch an den Ungenauigkeiten der Simulationen von etwa 3-4%.

5.5.4 Vergleich mit anderen messungsbasierten Verfahren

Zusätzlich zum Vergleich mit dem Peakrate-Ansatz aus [R01] ist der Vergleich mit anderen messungsbasierten Verfahren von Interesse. Abbildung 5.22 zeigt die GoS-Kurven des Time Window Verfahrens aus [JSDZ96] und des Point Sample Verfahrens aus [BSJ97]. Die beiden Verfahren wurden für diese Simulationen in die Handoverpriorisierung integriert, indem die Komponente FF_Estimator (siehe Abschnitt 5.3.1) durch Objekte ersetzt wurde, die die betreffenden Verfahren implementieren. Um die Ergebnisse mit der ebenfalls eingezeichneten GoS-Kurve (grau) von HoPVarB aus Abschnitt 5.5.1 vergleichen zu können, ist das Szenario bis auf das Modul zur Schätzung der Netzwerkauslastung das gleiche wie dort.

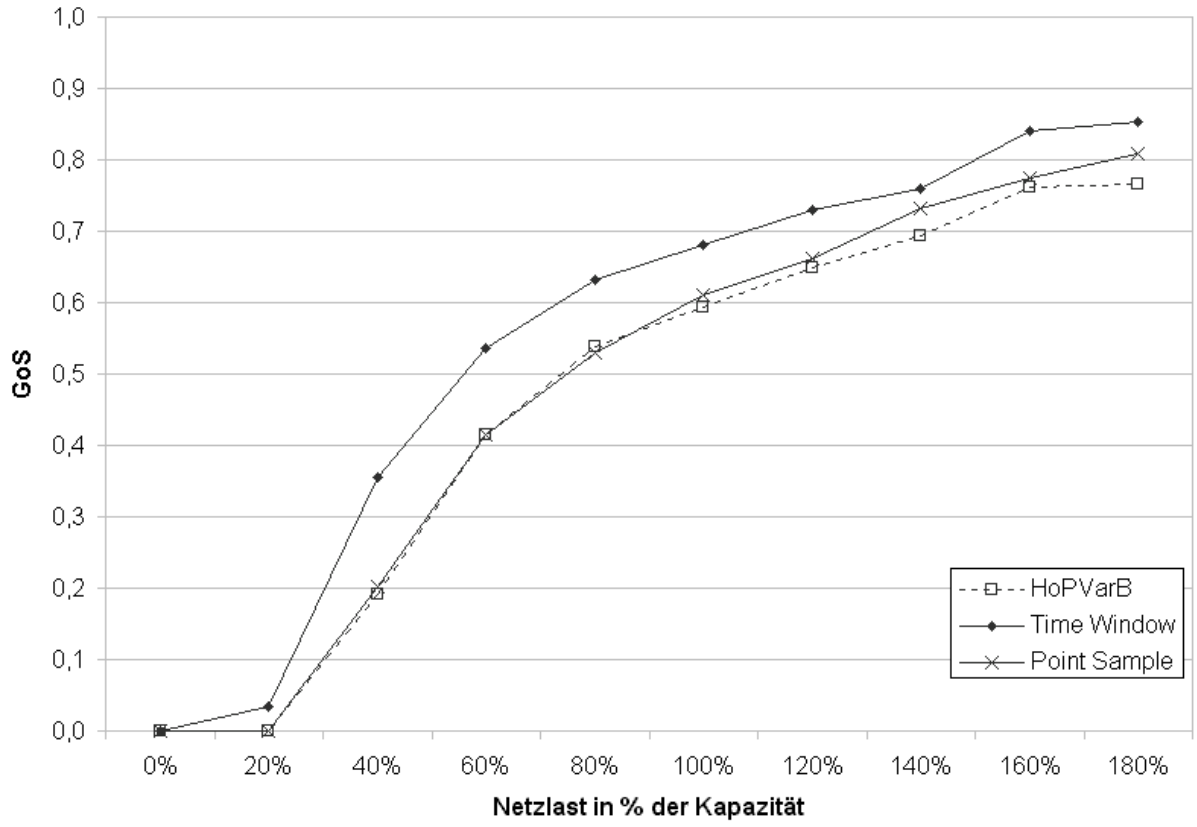

Abbildung 5.22: GoS-Kurven der Handoverpriorisierung mit verschiedenen MBAC

Es ist zu erkennen, dass die GoS-Kurve von HoPVarB am flachsten und somit am günstigsten verläuft. Während das Point Sample Verfahren erst mit zunehmender Last im Vergleich schlechter abschneidet, liegt die Kurve des Time Window Verfahrens bei jeder Netzlast deutlich über den beiden anderen Kurven.

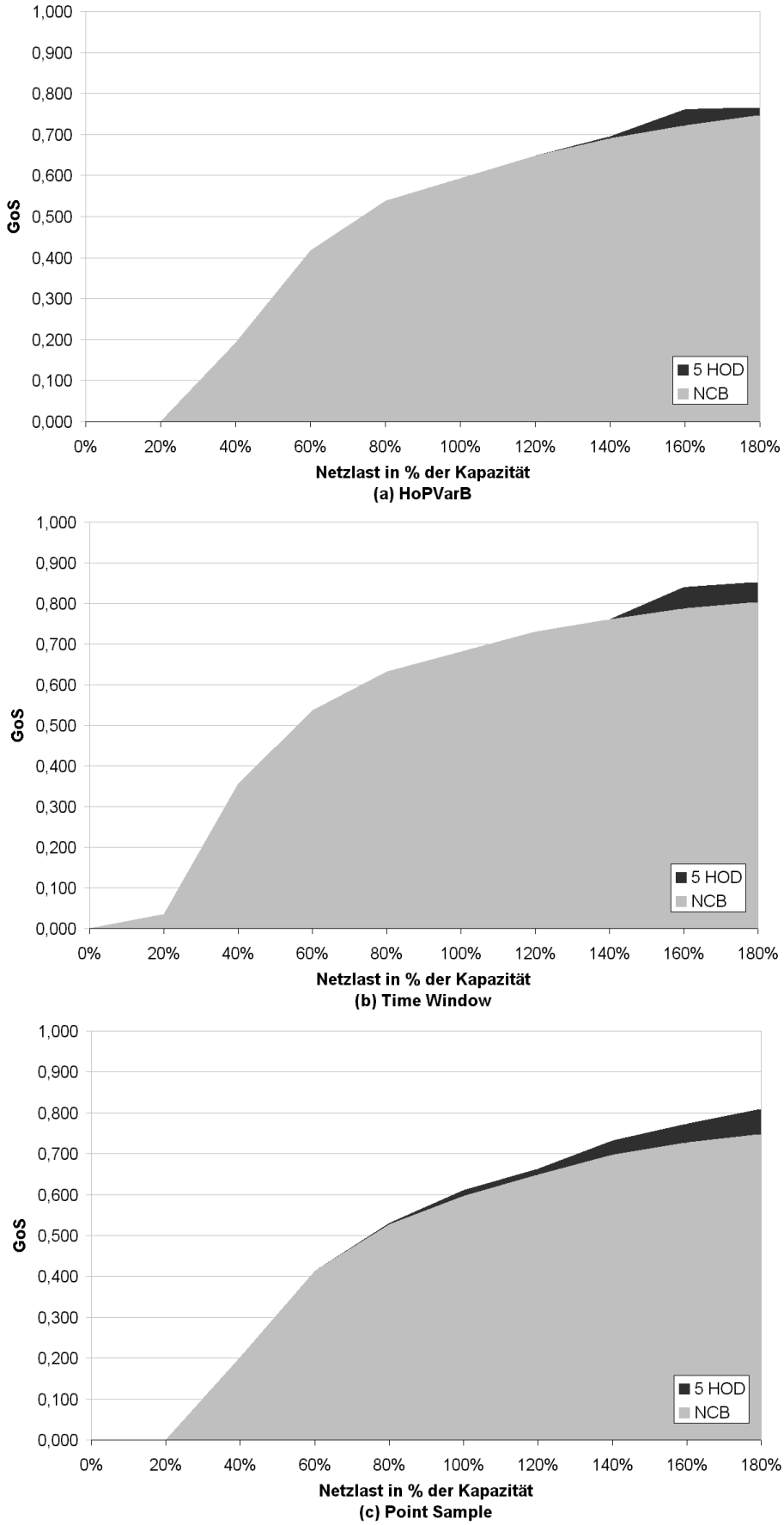

Abbildung 5.23: GoS-Kurven verschiedener MBACs differenziert nach NCB und HOD

Dieses Ergebnis ist wenig überraschend und entspricht den in Abschnitt 3.3.4 geäußerten Erwartungen über das Verhalten der Verfahren. Die Gründe dafür sind in Abbildung 5.23 erkennbar. Sie zeigt die GoS-Kurven der drei Verfahren, dabei sind der auf New Call Blocks zurückgehende Anteil (grau) und der auf Handover Drops zurückzuführende Anteil (schwarz) getrennt dargestellt. Es ist erkennbar, dass die Kurven der drei Verfahren unterschiedlich ausfallen.

Die Kurve der New Call Blocks von HoPVarB (a) verläuft flach, erste Handover Drops treten ab einer Last von 140% der Netzkapazität auf.

Bezüglich der Handover Drops ist die Situation bei dem Time Window Verfahren (b) ähnlich: auch hier treten erst bei 140% Drops auf. Allerdings ist ihre Zahl größer als bei HoPVarB. Anders ist die Situation bei den New Call Blocks: Die Kurve verläuft viel steiler und höher. Offensichtlich weist das Verfahren deutlich mehr neue Verbindungen ab als HoPVarB. Dies entspricht der Erwartung, dass der Algorithmus recht konservativ arbeitet. Tendenziell werden zu hohe Schätzwerte ermittelt. Das führt zu einer schlechteren Netzauslastung.

Umgekehrt ist die Situation bei dem Point Sample Verfahren (c). Der auf New Call Blocks zurückgehende Anteil der GoS-Kurve verläuft ähnlich flach wie bei HoPVarB. Es treten jedoch bereits ab einer Last von 80% die ersten Handover-Drops auf, ihr Anteil steigt dann mit zunehmender Last beständig an. Das weist darauf hin, dass das Verfahren die Netzlauslastung regelmäßig zu optimistisch einschätzt und in diesen Situationen zu viele Verbindungen zulässt. Andererseits verläuft der NCB-Anteil der GoS-Kurve nicht flacher als bei HoPVarB. Dies ist ein Hinweis darauf, dass die Auslastungsschätzung gelegentlich auch zu pessimistisch ist, und die Mehrzulassungen aufgrund überoptimistischer Schätzungen ausgleicht. Das unterstreicht die in Kapitel 3.3.4 geäußerte Erwartung, dass die stark schwankenden Messergebnisse die tatsächliche Auslastung regelmäßig sowohl über- als auch unterschätzen.

Das Equivalent Capacity Verfahren aus [F96] wurde für den Vergleich nicht herangezogen, da es für die Beschreibung neuer Datenströme anstelle der mittleren deren maximale Datenrate verwendet.

5.5.5 Einfluss des Verbindungsfaktors

In diesem Abschnitt soll geklärt werden, inwieweit die zur Berücksichtigung von beginnenden und endenden Verbindungen in den Schätzungen der Netzlast eingeführten Verbindungsfaktoren (siehe auch Kapitel 4.3.3) die Leistungsfähigkeit von HoPVarB positiv oder negativ beeinflussen.

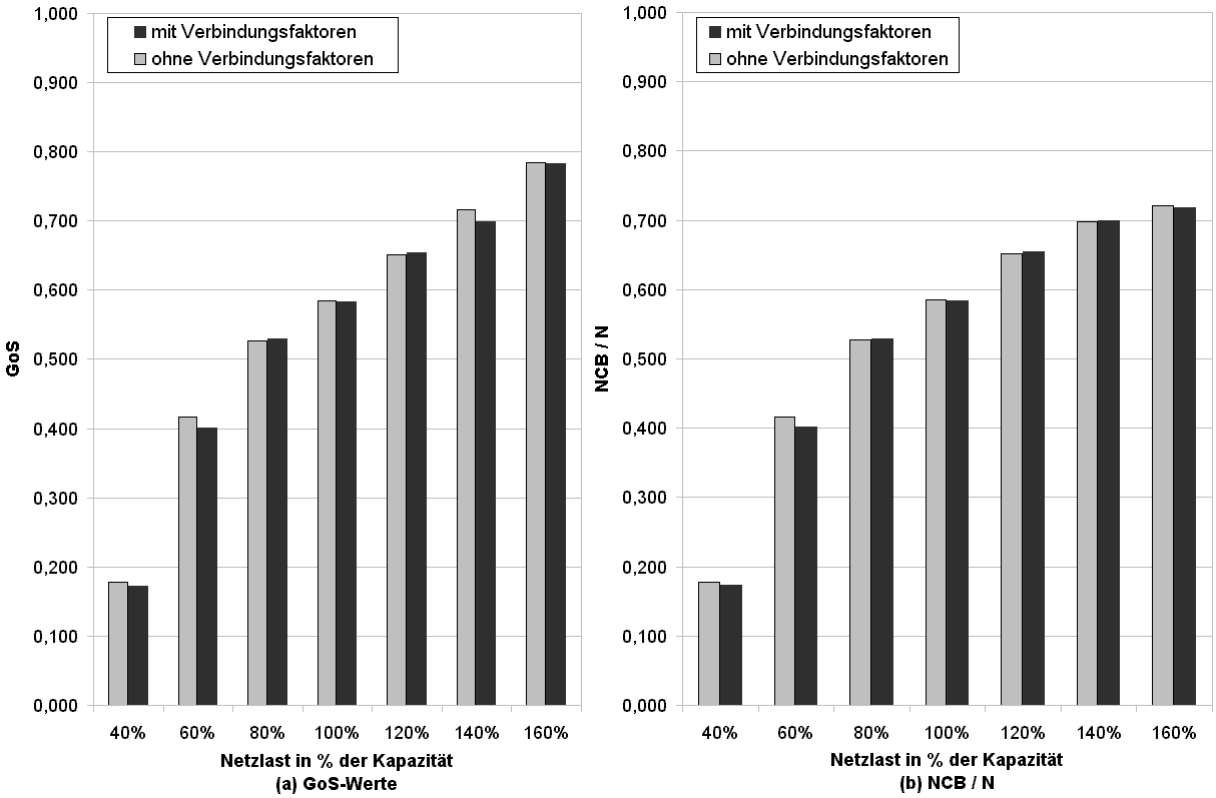

Abbildung 5.24: Einfluss des Verbindungsfaktors

Abbildung 5.24 (a) zeigt die GoS-Werte eines der beiden Simulationsläufe aus Kapitel 5.5.1 bei verschiedenen Netzlasten mit eingeschalteten (schwarz) und ausgeschalteten (grau) Verbindungsfaktoren im Vergleich.

Es ist erkennbar, dass der Einsatz der Verbindungsfaktoren in den meisten Fällen zu günstigeren (d.h. niedrigeren) GoS-Werten führt. Nur bei einer Last von 140% erzielt das Verfahren ohne Verbindungsfaktoren ein günstigeres Ergebnis. Insgesamt sind die Unterschiede zwischen den beiden Alternativen jedoch gering. Es ist interessant, dass sich keine klare Tendenz (wie etwa: bei hoher Last bringt der Einsatz der Verbindungsfaktoren mehr/weniger) ablesen lässt.

Abbildung 5.24 (b) zeigt die Anzahl der abgewiesenen Verbindungen New Call Blocks im Verhältnis zu der Gesamtzahl der Verbindungswünsche N. Das Diagramm lässt also im Vergleich mit (a) die Handover Drops außer Acht. Auch hier werden die Varianten mit und ohne Verbindungsfaktoren miteinander verglichen. Es wird deutlich, dass die Balken im Vergleich zueinander fast das gleiche Bild zeichnen wie in (a): Offensichtlich haben die Handover Drops fast keinen Einfluss darauf, welches der beiden Verfahren bei den verschiedenen Lastsituationen günstiger abschneidet. Lediglich bei einer Last von 140% gibt es eine geringe Abweichung von etwa 2%.

5.5.6 Selbstähnlicher Datenverkehr

Wesentliche Quellen von Datenverkehr mit variabler Bitrate sind Anwendungen, die komprimierte Audio- und/oder Videodaten versenden. Solche Datenströme haben häufig die Eigenschaft, selbstähnlich zu sein [GW94, F96]. Ein so gearteter Verkehr lässt sich durch die Überlagerung mehrerer Pareto-On-Off-Quellen (siehe Abschnitt 5.4.2) erzeugen [PF95]. Da Anwendungen mit variabler Bitrate im Mittelpunkt dieser Arbeit stehen, wird abschließend die Leistungsfähigkeit von HoPVarB hinsichtlich selbstähnlichen Verkehrs unter Beweis gestellt.

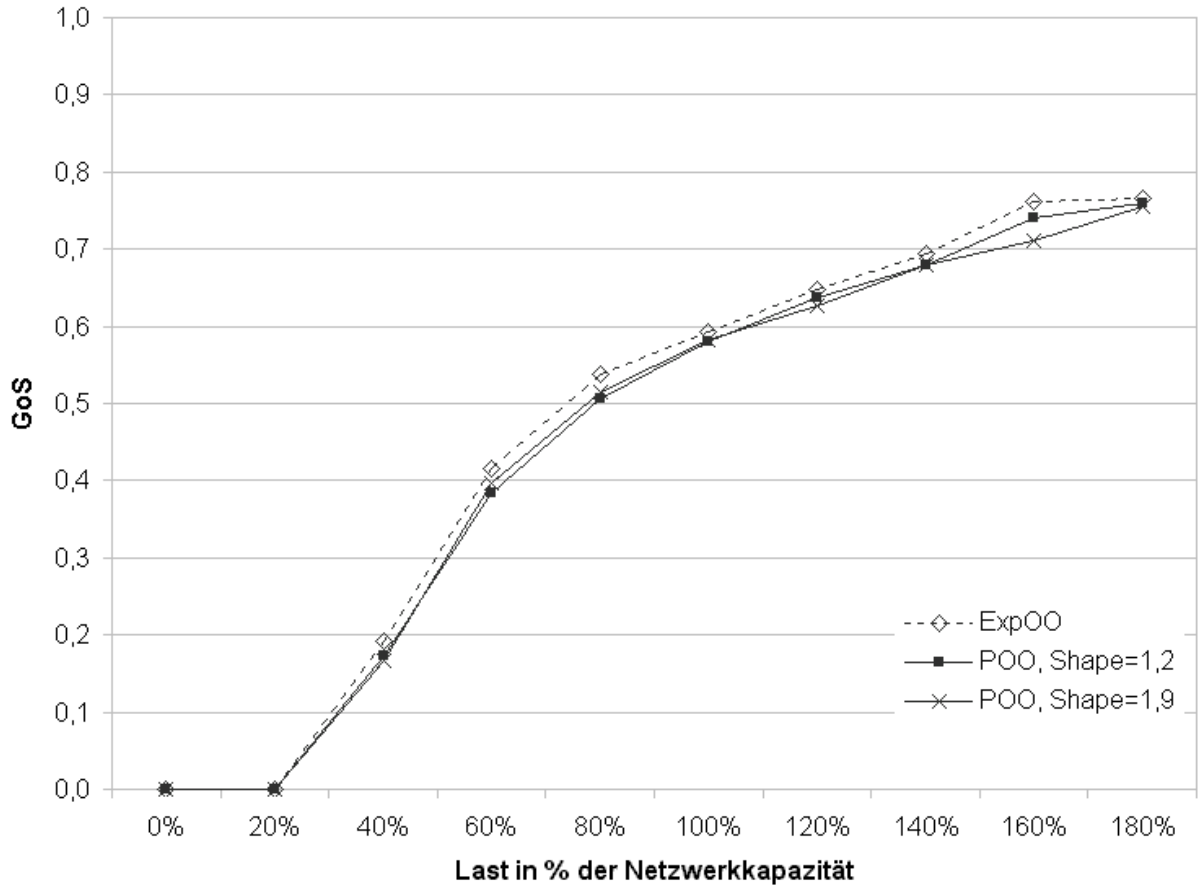

Abbildung 5.25: GoS-Kurve von Pareto- und Exponential-On-Off Quellen

Abbildung 5.25 zeigt den Verlauf der GoS-Kurven für selbstähnlichen Verkehr (schwarz). Zum Vergleich ist die Kurve aus dem in Kapitel 5.5.1 beschriebenen Szenario (grau) abgebildet. Um die Vergleichbarkeit sicherzustellen, sind die Szenarien nahezu gleich. Der einzige Unterschied ist, dass nicht exponentialverteilte, sondern paretoverteilte Datenquellen zum Einsatz kamen.

Es ist erkennbar, dass die Leistungsfähigkeit von HoPVarB auch bei selbstähnlichem Verkehr gegeben ist. Die beiden GoS-Kurven der paretoverteilten Verkehrsquellen liegen sehr nahe an der Kurve der exponentialverteilten Verkehrsquelle. Unabhängig von β sind die GoS-Werte

etwas niedriger als bei nicht selbstähnlichem Verkehr. Die Werte für New Call Blocks und Handover Drops weichen nicht nennenswert von denen in Abschnitt 5.5.1 ab.

Bemerkenswert ist, dass die Wahl des Shape-Parameters β nahezu keinen Einfluss auf den Verlauf der Kurven hat. Es wäre möglich gewesen, dass sich durchaus Unterschiede ergeben, da für β=1,2 die Streuung der On- bzw. Off-Zeiten größer ist. Die Abweichung zwischen den beiden Kurven liegt jedoch im Bereich der Simulationsgenauigkeit.

Insgesamt lässt sich somit feststellen, dass HoPVarB bei selbstähnlichem Verkehr die gleiche Leistungsfähigkeit zeigt wie bei Verkehrsmustern, die nicht selbstähnlich sind.

5.6 Zusammenfassung

In diesem Kapitel wurde die Leistungsfähigkeit von HoPVarB simulativ untersucht. Dazu wurden zunächst die verwendeten Modelle und der Simulator beschrieben. Im Zentrum der Betrachtungen standen die Simulationsergebnisse sowie ihre Bewertung mittels der Kostenfunktion Grade of Service (GoS).

HoPVarB begrenzt die Anzahl der Handoverdrops auch bei starker Netzüberlastung effektiv, indem nur so viele Verbindungen zugelassen werden, wie unterstützt werden können. Erst bei einer Last von 160% der Netzkapazität kommt es zu ersten Handover Drops; ihr Anteil liegt unter einem Prozent der begonnenen Verbindungen.

Mittels des Steuerparameters Reservierungshöhe lässt sich das Verhältnis von New Call Blocks und Handover Drops beeinflussen. Mit abnehmender Reservierungshöhe gehen die New Call Blocks zurück, während die Handover Drops zunehmen. Ein Reservierungsniveau kann entweder anhand der GoS-Funktion oder anhand der prozentualen Anzahl der Handover Drops gewählt werden.

In Vergleichen mit dem Peakrate-Verfahren wird die Effizienz von HoPVarB erkennbar: Bei Verkehr mit variabler Bitrate kann das Netzwerk deutlich besser ausgelastet werden, da HoPVarB statt der Peakrate die mittlere Datenrate der Verbindung für die Ressourcenbemessung zugrunde legt. Auch bei Datenverkehr mit konstanter Bitrate sind durch den Einsatz von HoPVarB keinerlei Nachteile gegenüber dem Peakrate-Verfahren hinzunehmen: Die GoS-Kurven liegen gleichauf. Darüber hinaus liefert HoPVarB im Vergleich mit anderen Verfahren zur messungsbasierten Schätzung der Netzwerklast ebenfalls die insgesamt besten Ergebnisse.

Abschließend wurde HoPVarB auf die Robustheit gegenüber selbstähnlichem Verkehr untersucht. Die Überlagerung von Verkehr aus paretoverteilten Quellen zeigt, dass die Ergebnisse marginal günstiger ausgefallen sind als bei exponentialverteilten Verkehrsquellen.

Insgesamt konnte gezeigt werden, dass HoPVarB in allen untersuchten Szenarien effizient arbeitet und gegenüber den anderen zum Vergleich herangezogenen Verfahren eine gute Netzlauslastung erzielt.

6 Zusammenfassung

In dieser Arbeit wurde das HoPVarB-Verfahren zur Handoverpriorisierung entwickelt. Grundlegendes Ziel war es dabei, die Wahrscheinlichkeit von Handover Drops durch die Reservierung von Ressourcen speziell für Handovers zu minimieren. Ein besonderes Augenmerk bei der Entwicklung lag dabei auf Datenströmen mit variabler Bitrate: Die effiziente Ermittlung der von den Strömen insgesamt erzeugten mittleren Datenrate mit Hilfe von Messungen der Netzwerkauslastung stand im Mittelpunkt.

Mit dem vorgestellten Verfahren ist es gelungen, alle gestellten Anforderungen zu erfüllen. Wichtigstes Kriterium ist eine gute Netzwerkauslastung. Durch die Messungen der Zellauslastung wird näherungsweise nur der Betrag an Bandbreite als belegt angesehen, den die Datenströme insgesamt zur Datenübertragung nutzen. Da auch die Reservierungen in den Nachbarzellen auf den Messungen beruhen und virtuell erfolgen, wird hier ebenfalls eine gute Effizienz erzielt.

Auch die Ziele hinsichtlich der Skalierbarkeit werden erreicht: Speicher- und Berechnungskomplexität sind von Nutzerzahl und Netzwerkgröße unabhängig. Es kommt eine dezentrale Datenhaltung zu Einsatz, Auslastungsmessungen und Reservierungen erfolgen aggregiert.

Das Verfahren ist einfach und transparent. Es gibt vier eindeutig abgegrenzte Komponenten und nur einen externen Steuerparameter: Die Reservierungshöhe H variiert das Verhältnis von New Call Blocks und Handover Drops. Der Algorithmus beruht nicht auf komplexen theoretischen Annahmen, die das Verständnis erschweren. Da nur einfache Berechnungen mit konstantem Aufwand durchgeführt werden, ist die Echtzeitberechenbarkeit sichergestellt.

Durch periodische Signalisierung ist das System robust gegen Ausfälle von Teilnehmern oder Paketverluste. Geringe Abweichungen zwischen beantragter und tatsächlicher Datenrate zugelassener Datenströme wirken sich nicht auf spätere Zulassungsentscheidungen aus. Einzelne Messfehler sind ebenfalls unproblematisch, schwankende Messwerte werden gut geglättet.

Da die mittlere Datenrate für die Zulassungsentscheidung zugrunde gelegt wird, ist die Fairness gegenüber sehr burst-artigen Strömen sichergestellt.

Das Verfahren wurde in Simulationen mit dem Network Simulator 2 auf Funktions- und Leistungsfähigkeit untersucht. Dabei konnte gezeigt werden, dass HoPVarB Handover Drops auch bei hoher Netzwerklast effektiv begrenzt. Mit Hilfe des Steuerparameters Reservierungshöhe lässt sich das Ausmaß der Priorisierung von Handovers gegenüber New Calls steuern. Im Vergleich sowohl mit dem Peakrate-Ansatz als auch mit anderen messungsbasierten Verfahren konnte eine gute Ressourcenauslastung belegt werden. Dabei ergaben sich keine nennenswerten Unterschiede zwischen selbstähnlichem Verkehr und Datenströmen, die diese Eigenschaft nicht aufweisen. Es wurde auch sichtbar, dass die zur

Berücksichtigung von beginnenden und endenden Verbindungen eingeführten Verbindungsfaktoren die Leistungsfähigkeit von HoPVarB in verschiedenen Situationen verbessern.

6.1 Ausblick

Allerdings sind die Ursachen für Verbesserungen und teilweise auch Verschlechterungen der Leistung von HoPVarB durch die Verbindungsfaktoren noch unzureichend geklärt. Es ist zu vermuten, dass die vom Mobilitätsgenerator erzeugten Bewegungsszenarien hier einen starken Einfluss ausüben: Während einige Szenarien eine Leistungssteigerung durch die Verbindungsfaktoren ausweisen, zeigen andere Szenarien mit identischen Parametern Verschlechterungen. Offensichtlich haben die Zufallsvariablen, die zu abweichenden Bewegungsmustern führen, stärkeren Einfluss als die Parameter (wie beispielsweise die Netzlast) des Szenarios. Diese unerwartete Tatsache sollte in weiteren Simulationen untersucht werden.

Denkbar ist auch, dass dieses Verhalten durch die gewählte räumliche Anordnung von Basisstationen oder Bewegungskorridoren zustande kommen. Auch um die Simulationsergebnisse insgesamt zu bestätigen, sollten weitere Simulationen mit umfangreicheren Szenarien und anderen Verläufen der Bewegungskorridore durchgeführt werden. Dies war im Rahmen der vorliegenden Arbeit aufgrund der langen Simulationszeiten nicht möglich.

In Abschnitt 5.5.2 wurde deutlich, dass bei unterschiedlicher Netzwerklast verschiedene Einstellungen des Steuerparameters Reservierungshöhe (H) zu optimalen Ergebnissen führen. Daher ist zu überlegen, ob eine automatische Einstellung von H in Abhängigkeit von der Last in HoPVarB integriert werden kann, ohne dass dadurch die Einfachheit und Transparenz des Verfahrens beeinträchtigt werden. Wird eine solche automatische Optimierung eingeführt, sollte deren Leistungsfähigkeit im Vergleich zu verschiedenen manuellen Einstellungen simulativ überprüft werden.

7 Literatur

[BBC+98] S. Blake, D. Black, M. Carlson, E. Davies, Z.Wang, W. Weiss: „An Architecture for Differenciated Services“, Request for Comments 2475 (Informational), Internet Engineering Task Force, Dezember 1998

[BCS94] R. Braden, D. Clark, S. Shenker: „Integrated Services in the Internet Architecture: an Overview“, Request for Comments 1633 (Informational), Internet Engineering Task Force, Juni 1994

[BJS99] L. Breslau, S. Jamin, S. Shenker: „Measurement-Based Admission Control: What is the Research Agenda?“, IEEE/IFIP IWQoS ’99, London, Mai 1999

[BJS00] L. Breslau, S. Jamin, S. Shenker: „Comments on the Performance of Measurement-Based Admission Control Algorithms“, Proceedings of the Conference on Computer Communications (IEEE Infocom) 2000, Tel Aviv, März 2000

[D97] R. Droms: "Dynamic Host Configuration Protocol", Request for Comments 2131 (Standards Track), Internet Engineering Task Force, März 1997.

[F96] S. Floyd: „Comments on Measurement-based Admissions Control for Controlled-Load Services“, Technical Report, Lawrence Berkley Laboratory, Juli 1996

[FH98] P. Ferguson, H. Huston: „Quality of Service – Delivering QoS on the Internet and in Corporate Networks“, Wiley Computer Publishing, 1998

[FL00] D. Figueiredo, B. Liu: On the Specification of NS and other Known On.Off Sources“, Computer Science Technical Report 00-25, Universtity of Massachusetts, Amherst, Juni 2000

[FV02] K. Fall, K. Varadhan (Herausgeber): „The ns Manual (formerly ns Notes and Documentation)“, März 2002, http://www.isi.edu/nsnam/ns/doc/ns_doc.ps.gz

[G02] D. Grossman: „New Terminology and Clarifications for Diffserv“, Request for Comments 3620 (Standards Track), Internet Engineering Task Force, April 2002

[GK97] R. Gibbens, R. Kelly: „Measurement-Based Connection Admission Control“, 15th International Teletraffic Congress Proceedings, Juni 1997

[GPP02] 3rd Generation Partnership Project: Technical Specification Group, Services and System Aspects: „QoS Concept and Architecture (Release 1999)“, Technical Report, Januar 2002

[GT97a] M. Grossglauer, D. Tse: „Measurement Based Admission Control: Analysis and Simulation“, Proceedings of the Conference on Computer Communications (IEEE Infocom) ’97, Kobe, April 1997.

[GT97b] M. Grossglauer, D. Tse: „A Framework for Robust Measurement-Based Admission Control“, Proceedings of ACM Sigcomm ’97, September 1997

[GT99] M. Grossglauer, D. Tse: „A Time-Scale Decomposition Approach to Measurement-Based Admission Control“, Proceedings of the Conference on Computer Communications (IEEE Infocom) ’99, S. 1539-1547, NewYork, März 1999

[GW94] M. Garret, W. Willinger: „Analysis, Modeling and Generation of Self-Similar VBR Video Traffic“, Proceedings of ACM SigComm, London, September 1994

[HBW+99] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski: „Assured Forwarding PHB Group“, Request for Comments 2597 (Standards Track), Internet Engineering Task Force, Juni 1999

[HR99] D. Hong, S. Rappapert: „Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures“, IEEE Transactions on Vehicular Technology, August 1986, siehe auch: CEAS Technical Report No 773, State University of New York, June 1999

[JDS+96] S. Jamin, P. Danzig, S. Shenker, L. Zhang: „A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks (Extended Version)“, ACM/IEEE Transaction on Networking, Dezember 1996

[JSD97] S. Jamin, S. Shenker, P. Danzig: „Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service”, Proceedings of the Conference on Computer Communications (IEEE Infocom) ’97, Kobe, April 1997.

[KN01] M. Kim, B. Noble: „Mobile Network Estimation“, Proceedings of the ACM Conference on Mobile Computing and Networking, Rome, Juni 2001

[NBB+98] K. Nichols, S. Blake, R. Baker, D. Black: „Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers“, Request for Comments 2474 (Standards Track), Internet Engineering Task Force, Dezember 1998

[NS2] The Network Simulator - ns-2, http://www.isi.edu/nsnam/ns/index.html

[O94] J.K. Ousterhout: „Tcl and the the TK toolkit“, Addison-Wesley, 1994

[P80] J. Postel: "User Datagram Protocol", STD 6, Request for Comments 768, Internet Engineering Task Force, August 1980.

[P81] J. Postel: „Internet Protocol“, Request for Comments 791, Internet Engineering Task Force, September 1981.

[P84] J. Postel: "Multi-LAN Address Resolution", Request for Comments 925, Internet Engineering Task Force, Oktober 1984.

[P96a] G. Pollini: „Trends in Handover Design“, IEEE Communications Magazine, Vol. 34, Nr. 3, März 1996

[P96b] C. Perkins: "IP Encapsulation within IP", Request for Comments 2003 (Standards Track), Internet Engineering Task Force, Oktober 1996.

[P02] C. Perkins: „IP Mobility Support IPv4“, Request for Comments 3220 (Standards Track), Internet Engineering Task Force, Januar 2002

[PF95] V. Paxson, S. Floyd: „Wide-Area Traffic: The Failure of Poisson Modeling“ IEEE/ACM Transactions on Networking, Vol. 3, No. 3, Juni 1995

[QK01] J. Qui, W. Knightly: „Measurement-Based Admission Control with Aggregate Traffic Envelopes”, IEEE/ACM Transactions on Networking, Vol. 9, No. 2, April 2001

[R00] M. Reislein: „Measurement-Based Admission Control for Bufferless Multiplexers“, Technical Report, GMD Fokus, Berlin, August 2000

[R01] L. Roth: „Variable Handover Priorisierung für Echtzeit-Anwendungen in mobilen DiffServ Netzen“, Diplomarbeit, Technische Universität Braunschweig, Juli 2001

[SS97] M. Sidi, D Starobinski: „New Call Blocking versus Handoff Blocking in Cellular Networks“, ACM Journal of Wireless Networks, Vol. 3, No. 1, März 1997

[S00] B. Stroustrup: „The C++ Programming Language“, 3. Auflage, Addison-Wesley, Februar 2000

[TUT] „Marc Greis’ Tutorial for the UCB/LBNL/VINT Network Simulator ns“, http://www.isi.edu/nsnam/ns/tutorial/index.html

[WDG96] U. Weiss, U. Dropmann, P. Godlewski: "Admission Control Guaranteeing Quality of Service in Packet Access Based Pico Cellular Networks", Proceedings of the 1996 ACTS Mobile Telecommunications Summit, Granada, S. 204-210, November 1996

[XN99] X. Xiao, L. Ni: „Internet QoS: A Big Picture“, IEEE Network, Vol. 13, No. 2, März 1999