

Universität Stuttgart

Institut für Nachrichtenvermittlung und Datenverarbeitung

Prof. Dr.-Ing. habil. Dr. h. c. mult. P. J. Kühn

77. Bericht über verkehrstheoretische Arbeiten

Mobilität in Kommunikationsnetzen

– Konzepte, Modellierung und Leistungsbewertung

von

Michael Schopp

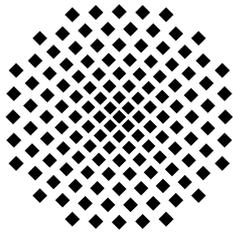
2001

D 93

© 2001 Institut für Nachrichtenvermittlung und Datenverarbeitung, Universität Stuttgart

Druck: E. Kurz & Co., Druckerei + Reprografie GmbH., Stuttgart

ISBN 3-922403-87-5



University of Stuttgart

Institute of Communication Networks and Computer Engineering

Prof. Dr.-Ing. habil. Dr. h. c. mult. P. J. Kühn

77th Report on Studies in Congestion Theory

**Mobility Support in Communication Networks
– Concepts, Modelling and Performance Evaluation**

by

Michael Schopp

2001

Summary

It is expected that a significant share of future information and communications services will be provided in a personalised way. Each subscriber will have access to such services anytime and anywhere. Supporting the mobility of persons, machines and terminals in a heterogeneous communications infrastructure will be one of the central building blocks of future service provisioning. To support mobility between networks and between network technologies for a large number of people requires scalable and powerful mobility management schemes. In this thesis, the phenomenon mobility is described in the context of the OSI reference model. The basic mechanisms and the different possibilities to support mobility in classical telecommunications networks as well as in IP-based networks are systematically analysed and classified. The fundamental trade-off between traffic routing, destination searching and data updating is identified. Based on these considerations, a number of new proposals how to advance existing mobility management schemes are presented. Three proposals are worth to be highlighted here: 1) Sending early binding updates in Mobile IP, 2) using Proxy HLRs in GSM/UMTS, and 3) establishing a common distributed database for the support of mobility between different types of networks.

In order to enable quantitative and qualitative comparisons between different mobility management schemes, suitable modelling of the behaviour of mobile instances is required. In this thesis, a model of the behaviour of mobile instances is proposed which does not only model the mobility of the instances but also the processes which initiate the search of a mobile instance (i.e. 'incoming calls'). Important aspects of the work are the development of efficient and effective methods for the parameterisation of the mobility model, and the introduction of different 'types of calls' which describe origin and destination of calls (i.e. the traffic matrix) in relation to the mobility behaviour of the involved mobile instances. Together with an appropriate load model, the model of the behaviour of mobile instances is used to create model traffic which can be used to study the performance of different mobility management schemes. The transport of signalling messages and the execution of mobility management related processes in a distributed system is modelled by using a flexible network model. This network model is designed to enable a rather simple analytical performance evaluation even for the case where the serving stations are modelled in a rather complex way.

The thesis is completed by comparing the existing GSM mobility management with schemes which could potentially be used in GSM or in UMTS and which are either enhancements of the existing mobility management scheme or alternative solutions. The comparison is based on two criteria which both depend on the parameters of the behaviour of the mobile instances: a) the network load which is generated by the identified model traffic, and b) the expected

response times of different mobility management procedures. In all cases, the parameters are chosen out of a wide range of possible values. Therefore, the results provide answers regarding the performance of different mobility management schemes for a broad range of application scenarios. Mobile documents in the World Wide Web usually exhibit a different behaviour than mobile agents or mobile computational objects. Those exhibit a different behaviour than UPT-subscribers, the destinations of freephone numbers (0800), or portable telephone numbers. Some mobility management schemes are only suitable for a small part of the parameter space but may exhibit significant advantages for specific applications over other schemes which may be suitable for a broader range of applications.

The core of the thesis is composed of two parts. Part one consists of the chapters 2 to 4. Part two consists of chapter 5 and chapter 6. Both parts are arranged in a similar form: First, fundamental considerations and the basic methodology are introduced. Then, the obtained insight into the principal behaviour and the proposed methods are applied to some selected application scenarios.

Chapter 2 describes mobility in the context of the OSI reference model and outlines the fundamental network technologies (SS7, Intelligent Networks, mobility management in GSM, Mobile IP). Chapter 3 defines the tasks of mobility management and classifies different mobility management schemes. Important aspects are the interaction between traffic routing and destination searching, and the different options for placing, searching and updating of mobility related data in the network. In Chapter 4 several proposals for advanced mobility management schemes are made which can be classified in three groups: 1) schemes which support optimal routing of user traffic, 2) schemes which advance the existing GSM mobility management, and 3) proposal for a common mobility management which spans different network technologies.

In Chapter 5, the model of the behaviour of mobile instances, the load model, and the network model are described. An important aspect of the chapter is the parameterisation of the mobility model. Chapter 6 compares different mobility management schemes with respect to some important criteria. The thesis is concluded by a short summary and an outlook.

Inhaltsverzeichnis

Summary	i
Inhaltsverzeichnis	iii
Abkürzungen	vii
Formelzeichen	xiii
1 Einleitung	1
1.1 Mobilitätsunterstützung als Grundlage für personenbezogene Dienste	1
1.2 Übersicht über die Arbeit	4
2 Netztechnische Grundlagen	7
2.1 Adressierung, Numerierung und Mobilität	8
2.1.1 Adressierung und Mobilität bei OSI	8
2.1.2 Adressierung und Mobilität im Internet	12
2.1.3 Numerierung und Adressierung in klassischen Telekommunikationsnetzen	14
2.2 Das Signalisiersystem Nummer 7	15
2.3 GSM und UMTS	18
2.3.1 GSM – Global System for Mobile Communications	18
2.3.2 Teilnehmerlokalisierung in GSM	22
2.3.3 UMTS – Universal Mobile Telecommunication System	25
2.4 Das Konzept des „Intelligenten Netzes“	26
2.4.1 Das konzeptionelle Modell des „Intelligenten Netzes“	28
2.4.2 Internationale Standardisierung des „Intelligenten Netzes“	31
2.4.3 TINA	39
2.4.4 Mobilitätsverwaltung im Rahmen des „Intelligenten Netzes“	43
2.5 Mobilität in IP-basierten Netzen: <i>Mobile IP</i>	47
2.5.1 Die Funktionsweise von <i>Mobile IP</i>	47
2.5.2 Die weitere Entwicklung von <i>Mobile IP</i>	50
3 Mobilitätsverwaltung	53
3.1 Aufgaben der Mobilitätsverwaltung	53
3.2 Zusammenspiel von Wegführung und Zielsuche	55
3.3 Datenhaltung	58
3.3.1 Plazierung von Daten	61
3.3.2 Ausfindigmachen von Daten	63
3.3.3 Datenaktualisierung	65
3.4 Verbindungsweitergabe	68

3.5	Beispiele	69
3.5.1	Datenhaltung in GSM	69
3.5.2	Hierarchische Datenhaltung für universelle Mobilität	71
3.5.3	Hierarchisches Datenverzeichnis in MONET	74
3.5.4	<i>Mobile IP</i> mit <i>Route Optimization</i>	78
3.5.5	Die <i>Datacycle Architecture</i>	80
3.5.6	<i>Mobile PNNI</i>	81
3.5.7	Erweiterungen der vorgestellten Verfahren	82
4	Vorschläge zur Weiterentwicklung von Verfahren der Mobilitätsverwaltung	85
4.1	Unterstützung von optimaler Nutzverkehrsführung	85
4.1.1	Mobilitätsunterstützung im PSTN/ISDN	86
4.1.2	Auslösen der Zielsuche bei gemeinsamem Numerierungsraum	87
4.1.3	Frühes <i>Binding Update</i> für <i>Mobile IP</i>	88
4.2	Fortentwicklung der GSM-Mobilitätsverwaltung	93
4.2.1	Zusätzliche Datenkopien	94
4.2.2	Einsatz der <i>Datacycle Architecture</i>	95
4.2.3	Verzicht auf Ortsaktualisierungen im HLR	97
4.2.4	Einführung von <i>Proxy HLRs</i>	98
4.3	Netztechnikenübergreifende Mobilitätsunterstützung	100
4.3.1	Zusammenwirken von UPT und GSM	101
4.3.2	Netzeübergreifende Aufenthaltsortsverwaltung	102
4.4	Zwischenbilanz der Arbeit	103
5	Modellierung	105
5.1	Das Teilnehmermodell: Bewegung	106
5.1.1	Grundlagen	107
5.1.2	Stationäres Modell	111
5.1.3	Parametrisierung des stationären Modells	113
5.1.4	Überlagerung von Bewegungsmodellen	120
5.1.5	Kopplung des Bewegungsverhaltens und transientes Modell	122
5.2	Das Teilnehmermodell: Rufe	124
5.2.1	Darstellung als Verkehrsmatrix	125
5.2.2	Verwendung von Rufotypen	126
5.2.3	Zusammenfassung der Vorgehensweise	131
5.3	Das Lastmodell	135
5.3.1	Prinzipielle Vorgehensweise	135
5.3.2	Angepaßte Vorgehensweise	136
5.3.3	Zusätzliche Beschreibungsmittel	138
5.4	Das Netzmodell	142
5.4.1	Grundlegende Modellkomponenten	142
5.4.2	Modellierung der Netzknoten	143
5.5	Zusammenfassung	147
6	Leistungsbewertung ausgewählter Beispiele	149
6.1	Voraussetzungen	150
6.1.1	Untersuchte Netzstruktur und Parameter des Teilnehmermodells	150
6.1.2	Die untersuchten Verfahren zur Aufenthaltsortsdatenverwaltung	151

6.2	Vergleich auf der Basis des Verkehrsangebots	153
6.2.1	Plazierung von Netzknoten bei GSM-basierter Mobilitätsverwaltung	153
6.2.2	Vergleich mit alternativen Verfahren	155
6.2.3	Weitere Einflüsse auf die Kosten einer Teilnehmersuche	160
6.2.4	Zahl der Datenbankeinträge und Datenbankzugriffe	163
6.2.5	Verkehrsangebot je Netzinstanz	166
6.3	Vergleich auf der Basis von Antwortzeiten	167
6.3.1	Methodik zur Bestimmung von Antwortzeiten	168
6.3.2	Parameter des Netzmodells und Bestimmung von Antwortzeiten	170
6.3.3	Ergebnisse	171
6.4	Fazit	174
7	Zusammenfassung und Ausblick	175
	Literaturverzeichnis	179
	Anhang A	191
A.1	Erweitertes Beispiel zur Parametrisierung des stationären Zustandsmodells	191
A.2	Weitere Ergebnisse der Leistungsbewertung aus Kapitel 6	192
	Anhang B	199
B.1	Das SSF/CCF Modell im IN CS-1	199
B.2	Die zugehörigen <i>Basic Call State Models</i>	200

Abkürzungen

ACM	Address Complete Message (ISUP, SS7)
ACSE	Association Control Service Element (OSI)
AIN	Advanced Intelligent Network (Bellcore)
AMPS	Advanced Mobile Phone System
ANM	Answer Message (ISUP, SS7)
APDU	Application Protocol Data Unit (OSI)
ARP	Address Resolution Protocol (Internet)
as-UAP	access session related UAP (TINA)
ASE	Application Service Element (OSI)
ATM	Asynchronous Transfer Mode
B-ISDN	Breitband-ISDN
BCAF	Bearer Control Agent Function (IN/UMTS)
BCF	Bearer Control Function (IN/UMTS)
BCP	Basic Call Process (IN)
BCSM	Basic Call State Model (IN)
BCUP	Basic Call-Unrelated Process (IN)
Bellcore	Bell Communications Research
BSC	Base Station Controller (GSM)
BSS	Base Station System (GSM)
BTS	Base Transceiver Station (GSM)
CAMEL	Customized Applications for Mobile network Enhanced Logic (GSM)
CCAF	Call Control Agent Function (IN)
CCF	Call Control Function (IN)
CEPT	Conférence Européenne des Postes et Télécommunications
CLASS	Custom Local Area Signalling Service
CMR	Call-to-Mobility Ratio
CORBA	Common Object Request Broker Architecture (OMG)
CPE	Customer Premises Equipment
CRACF	Call-Related Radio Access Control Function (IN)
CS- <i>x</i>	Capability Set <i>x</i> (IN)
CSM	Communication Session Manager (TINA)
CURACF	Call-Unrelated Radio Access Control Function (IN)
CUSF	Call-Unrelated Service Function (IN)

CUSP	Call-Unrelated Service Point (IN)
DCS	Digital Cellular System (GSM)
DDB	Distributed Database (MONET)
DECT	Digital European Cordless Telecommunications (ETSI)
DFP	Distributed Functional Plane (IN)
DIS	Draft International Standard (ISO)
DNS	Domain Name System (Internet)
DPE	Distributed Processing Environment
DSL	Digital Subscriber Line
DSS 1	Digital Subscriber Signalling System No. 1 (ISDN)
DTMF	Dual Tone Multi Frequency
EIA	Electronic Industry Association
ETSI	European Telecommunications Standards Institute
FE	Functional Entity (IN)
FEA	Functional Entity Action (IN)
FPLMTS	Future Public Land Mobile Telecommunication System (heute: IMT-2000)
GFP	Global Functional Plane (IN)
GMM	Global Multimedia Mobility (ETSI)
GMSC	Gateway MSC (GSM)
GPRS	General Packet Radio Service (GSM)
GSM	Global System for Mobile Communications (Groupe Spécial Mobile) (ETSI)
GSM-R	GSM-Rail (GSM)
GTT	Global Title Translation (SS7)
HDB	Ortsdatenverwaltung mit einem hierarchischen, verteilten Datenbanksystem
HLR	Home Location Register (GSM)
HLSIB	High Level SIB (IN)
HSCSD	High Speed Circuit Switched Data (GSM)
IA	Initial Agent (TINA)
IAF	Intelligent Access Function (IN)
IAM	Initial Address Message (ISUP, SS7)
ICMP	Internet Control Message Protocol (Internet)
IDL	Interface Definition Language (OMG)
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force (Internet)
IMSI	International Mobile Subscriber Identity (GSM)
IMT-2000	International Mobile Telecommunications 2000 (früher: FPLMTS)(ITU-T)
IN	Intelligent Network
INA	Information Networking Architecture (Bellcore)
INAP	Intelligent Network Application Part (IN)

INCM	Intelligent Network Conceptual Model (IN)
IP	Intelligent Peripheral (IN)
IP	Internet Protocol (Internet)
IS-41	EIA/TIA Interim Standard 41
IS-54	EIA/TIA Interim Standard 54 („TDMA“)
IS-95	EIA/TIA Interim Standard 95 („CDMA“)
ISDN	Integrated Services Digital Network
ISO	International Organization for Standardization
ISUP	ISDN User Part (SS7)
IT	Information Technology, Informationstechnik
ITG	Informationstechnische Gesellschaft im VDE
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector (ehem. CCITT)
LAI	Location Area Identity (GSM)
LAN	Local Area Network
LAPD	Link Access Procedure on the D-channel (ISDN)
LCMR	Local Call-to-Mobility Ratio
MAP	Mobile Application Part (GSM)
MBCF	Mobile Bearer Control Function (IN/UMTS)
MCCF	Mobile Call Control Function (IN/UMTS)
MCF	Mobile Control Function (IN/UMTS)
MIX	Kunstwort aus dem englischen Verb <i>to mix</i> . Es ist kein Akronym.
MONET	RACE-Projekt „MOBILE NETWORKS“
MoU	Memorandum of Understanding
MS	Mobile Station (GSM)
MSC	Message Sequence Chart (Z.120)
MSC	Mobile Switching Centre (GSM)
MSF	Mobile Storage Function (IN/UMTS)
MSIN	Mobile Subscriber Identification Number (GSM, E.212)
MSISDN	Mobile Subscriber ISDN Number (GSM) nach GSM 01.04: Mobile Station ISDN Number
MSRN	Mobile Station Roaming Number (GSM)
MTP 3b	Ebene 3 des MTP bei B-ISDN (SS7)
MTP	Message Transfer Part (SS7)
NNI	Network Node Interface
ODL	Object Definition Language (TINA)
ODP	Open Distributed Processing
OMG	Object Management Group
OMSS	Operation and Maintenance Subsystem (GSM)

OSI	Open System Interconnection (ISO)
PA	Provider Agent (TINA)
PCS	Personal Communication Support (TINA)
PCS	Personal Communications System (GSM)
PCS-UA	PCS-enhanced User Agent (TINA)
PE	Physical Entity (IN)
PEA	Physical Entity Action (selbstgeprägter Begriff in Anlehnung an FEA)
PHLR	Proxy HLR
PLMN	Public Land Mobile Network
PNNI	Private Network Node Interface (ATM Forum)
POI	Point of Initiation (IN)
POR	Point of Return (IN)
PSTN	Public Switched Telephone Network
RACE	Research and development in Advanced Communications technologies in Europe
RACF	Radio Associated Control Function (IN/UMTS)
RBCF	Radio Bearer Control Function (IN/UMTS)
RCF	Radio Control Function (IN)
RFC	Request for Comments (Internet)
ROSE	Remote Operation Service Element (OSI)
SAAL	Signalling ATM Adaptation Layer (ATM)
SAP	Service Access Point (OSI)
SCAF	Service Control Agent Function (IN/UMTS)
SCCP	Signalling Connection Control Part (SS7)
SCEF	Service Creation Environment Function (IN)
SCF	Service Control Function (IN)
SCP	Service Control Point (IN)
SCUAF	Service Control User Agent Function (IN)
SDF	Service Data Function (IN)
SDP	Service Data Point (IN)
SF	Service Factory (TINA)
SF	Service Feature (IN)
SIB	Service Independent Building Block (IN)
SIM	Subscriber Identity Module (GSM)
SMAF	Service Management Access Function (IN)
SMF	Service Management Function (IN)
SMS	Short Message Service (GSM)
SMSS	Switching and Management Subsystem (GSM)
SOR	Support of Optimal Routeing (GSM)

SRBCF	Specialized Resource Bearer Control Function (IN/UMTS)
SRF	Specialized Resource Function (IN)
ss-UAP	service session related UAP (TINA)
SS7	Signalling System No. 7
SSF	Service Switching Function (IN)
SSM	Service Session Manager (TINA)
SSN	Subsystem Number (SS7)
SSP	Service Switching Point (IN)
TACAF	Terminal Access Control Agent Function (IN/UMTS)
TC	Transaction Capabilities (SS7)
TCAP	Transaction Capabilities Application Part (Synonym für TC)
TCP	Transmission Control Protocol (Internet)
TCSM	Terminal Communication Session Manager (TINA)
TETRA	Trans European Trunked Radio System (ETSI)
TIA	Telephone Industry Association
TINA	Telecommunications Information Networking Architecture
TINA-C	TINA-Consortium
TMN	Telecommunication Management Network
TMSI	Temporary Mobile Subscriber Identity (GSM)
UA	User Agent (TINA)
UAP	User Application (TINA)
UDP	User Datagram Protocol (Internet)
UIC	Union Internationale des Chemins de Fer
UMTS	Universal Mobile Telecommunication System (ETSI)
UNI	User-Network Interface
UPT	Universal Personal Telecommunication
USM	User Service Session Manager (TINA)
VLR	Visitor Location Register (GSM)
VPN	Virtual Private Network
WWW	World Wide Web (Internet)
xDSL	<i>x</i> -beliebige DSL-Technik (z.B. ADSL)
X.500	OSI-Verzeichnissystem (ITU-T Empfehlungen der Serie X.500)

Formelzeichen

A	Teilmenge des Zustandsraums S
${}^{(i)}C_v^{(k)}$	Normierungsfaktor für Rufe vom Rufotyp i ($i \in \{I, II, III, IV\}$) zum Zustand v für Teilnehmer der Klasse k
$d_{u,v}$	Distanz zwischen Zustand u und Zustand v (u.U. auch: $d_{u,v}^{(k)}$)
δ	Spreizfaktor für die Gewichtung von Distanzen, $\delta \in [0, 1]$
δ_B bzw. δ_R	δ zur Beschreibung des Bewegungsmodells bzw. des Rufmodells
Δ	Symmetrische Distanzmatrix mit den Elementen $d_{u,v}$ (u.U. auch: $\Delta^{(k)}$)
\underline{e}	Einheitsvektor, bestehend aus Einsen
$\eta_u^{(k)}$	„Heimatafaktor“ des Zustands für Teilnehmer der Klasse k
$f_{\delta}(d)$	Gewichtungsfunktion zur Abbildung von Distanzen auf Gewichtungsfaktoren
$f_{u,v}(t)$	Verteilungsdichtefunktion der Verweildauer im Zustand u bevor ein Übergang in Zustand v erfolgt
$F_{u,v}(t)$	Verteilungsfunktion der Verweildauer im Zustand u bevor ein Übergang in Zustand v erfolgt
$F_{X_1 \dots X_n}$	n -dimensionale Verteilungsfunktion eines stochastischen Prozesses ($F_{X_1 \dots X_n}(x_1, \dots, x_n, t_1, \dots, t_n) = P\{X(t_1) \leq x_1, \dots, X(t_n) \leq x_n\}$).
$\Sigma\phi$ bzw. $\Sigma\phi^{(k)}$	Summe der globalen stationären Zustandsübergangsraten ϕ_u bzw. $\phi_u^{(k)}$
ϕ_u bzw. $\phi_u^{(k)}$	globale stationäre Zustandsübergangsrate („Fluß“) aus dem Zustand u heraus (globale Verlassensrate) (für Teilnehmer der Klasse k)
$\underline{\phi}$ bzw. $\underline{\phi}^{(k)}$	Vektor der globalen Verlassensraten ϕ_u (für Teilnehmer der Klasse k)
ϕ_u^*	mit Hilfe von Gewichtswerten G_u umgewichtete globale Verlassensrate ϕ_u
$\underline{\phi}^*$	Vektor der ϕ_u^*
Φ_u	globaler Fluß aller Teilnehmer aller Klassen aus Zustand u heraus
$\Phi_u^{(k)}$	globaler Fluß aller Teilnehmer der Klasse k aus Zustand u heraus
$\Phi_{u,v}^{(k)}$	globaler Fluß aller Teilnehmer der Klasse k von Zustand u zu Zustand v
$\phi_{u,v}$ bzw. $\phi_{u,v}^{(k)}$	globale stationäre Zustandsübergangsrate („Fluß“) zwischen den Zuständen u und v (für Teilnehmer der Klasse k)
$\Phi_{u,v}$	globaler Fluß aller Teilnehmer aller Klassen von Zustand u zu Zustand v
\mathbf{G}	Matrix mit den Elementen $g_{u,v}$
G_u	Gewichtswert, der einem Zustand u zugeordnet wird
$g_{u,v}$	Element der Matrix \mathbf{G}

$g_{u,v}(x)$	Verteilungsdichte der Gruppengröße beim Übergang von Zustand u zu Zustand v
$\Gamma_{u,v}$	mittlere Gruppengröße beim Übergang von Zustand u zu Zustand v
$h^{(k)}$	Heimatzustand eines Teilnehmers der Klasse k
$h, h_{\text{Suche}}, h_{\text{LUs}}$	Mittlere Bearbeitungszeiten in Datenbankknoten für die Operationen Teilnehmersuche und ds
i	allgemein verwendeter Index zur Bezeichnung eines Elements einer Menge
I	Einheitsmatrix – Matrix deren Hauptdiagonale aus Einsen besteht und deren restlichen Elemente Nullen sind.
k	Bezeichner für eine Teilnehmerklasse k
K	Menge aller Teilnehmerklassen
λ_i	Eigenwert einer Matrix
$\lambda_{u,v}^{(k)}$	Element von $\Lambda^{(k)}$, das die globale stationäre Rufrate zu einem Teilnehmer der Klasse k von u nach v angibt
λ_{\max}	Maximaler Eigenwert einer unzerlegbaren Matrix
$\Lambda^{(k)}$	globale, stationäre Verkehrsmatrix für Rufe zu einem Teilnehmer der Klasse k
M	beliebige Matrix
μ_u bzw. $\mu_u^{(k)}$	Kehrwert der mittleren Verweildauer im Zustand u (für Teilnehmer der Klasse k)
$\underline{\mu}$ bzw. $\underline{\mu}^{(k)}$	Vektor der mittleren Verweildauern μ_u (für Teilnehmer der Klasse k)
$\mu_{u,v}$ bzw. $\mu_{u,v}^{(k)}$	Kehrwert der mittleren Verweildauer im Zustand u bevor ein Übergang in Zustand v erfolgt (für Teilnehmer der Klasse k)
$n_{u,v}$	Element der Nachbarschaftsmatrix N , das angibt, ob zwischen Zustand u und Zustand v ein direkter Übergang zulässig ist
N	Nachbarschaftsmatrix mit den Elementen $n_{u,v}$
N	Zahl der Zustände des Zustandsraums S
π_u bzw. $\pi_u^{(k)}$	Zustandswahrscheinlichkeit für den Zustand u (für Teilnehmer der Klasse k)
$\underline{\pi}$ bzw. $\underline{\pi}^{(k)}$	Vektor der Zustandswahrscheinlichkeiten π_u (für Teilnehmer der Klasse k)
$\Pi^{(k)}$	Zahl der Teilnehmer der Teilnehmerklasse k
Π_u	mittlere Zahl von Teilnehmern im Zustand u
$\Pi_u^{(k)}$	mittlere Zahl von Teilnehmern der Klasse k im Zustand u
$\underline{\Pi}^{(k)}$	Vektor der $\Pi_u^{(k)}$
$p_{u,v}$ bzw. $p_{u,v}^{(k)}$	bedingte Übergangswahrscheinlichkeit von u nach v bei Zustandsprozessen (für Teilnehmer der Klasse k)
P bzw. P ^(k)	Zustandübergangsmatrix mit den Elementen $p_{u,v}$ (für Teilnehmer der Klasse k)
${}^{(i)}P_{\text{Ruf}}$ bzw. ${}^{(i)}P_{\text{Ruf}}^{(k)}$	Ruftypanteil des Ruftyps i mit $i \in \{I, II, III, IV\}$ (für einen Teilnehmer der Klasse k)

$Q_{u,v}(t)$	Hilfsfunktion zur Definition von $f_{u,v}(t)$ und $F_{u,v}(t)$ bei Semi-Markoffschen Zustandsprozessen (Verteilungsfunktionen der bedingten Aufenthaltsdauern im Zustand u)
ρ_{LU} s bzw. ρ_{Suche}	Verkehrsangebot an einem Prozessor eines Datenbankknotens, das von der Prozedur <i>Location Update</i> bzw. Teilnehmersuche herrührt
$r_v^{(k)}$	bedingte Rufraten zu einem Teilnehmer der Klasse k , wenn er sich in v befindet
$r_{u,v}^{(k)}$	bedingte Rufraten zu einem Teilnehmer der Klasse k vom „Ort“ u , wenn der gerufene Teilnehmer sich in v befindet
${}^{(i)}r_{u,v}$ bzw. ${}^{(i)}r_{u,v}^{(k)}$	Komponenten von ${}^{(i)}\mathbf{R}$ bzw. ${}^{(i)}\mathbf{R}^{(k)}$ ($i \in \{I, II, III, IV\}$) (für einen Teilnehmer der Klasse k)
Σ_r bzw. $\Sigma_r^{(k)}$	globale stationäre Rufrate (für einen Teilnehmer der Klasse k)
\mathbf{R} bzw. $\mathbf{R}^{(k)}$	Matrix der bedingten Rufraten $r_{u,v}$ bzw. $r_{u,v}^{(k)}$
${}^{(i)}\mathbf{R}$ bzw. ${}^{(i)}\mathbf{R}^{(k)}$	Bedingte Rufratenmatrix zum Ruftyp i mit $i \in \{I, II, III, IV\}$ (für einen Teilnehmer der Klasse k)
RBZ	Mittlere Restbedienzeit eines Prozessors eines Datenbankknotens
s	Element des kontinuierlichen Parameterraums T
s_{Suche} bzw. s_{LU} s	Mittlere Antwortzeit für eine Operation (Teilnehmersuche bzw. <i>Location Update</i>) an einem Datenbankknoten
S	Zustandsraum
t	Element des kontinuierlichen Parameterraums T
t_i	beliebiges Element des kontinuierlichen Parameterraums T
T	kontinuierlicher Parameterraum „Zeit“
T_n	Zeitpunkt des n -ten Zustandsübergangs
u	Bezeichner für einen Zustand des Zustandsraums S (auch: „Ort“ eines Teilnehmers)
v	Bezeichner für einen Zustand des Zustandsraums S (auch: „Ort“ eines Teilnehmers)
$X(t)$	Zustand eines stochastischen Prozesses zum Zeitpunkt t ($X \in S$)
$X(t, \xi_i)$	Zustand ($X \in S$) eines stochastischen Prozesses zum Zeitpunkt t im Zufallsexperiment ξ_i
ξ_i	i -tes Zufallsexperiment

Kapitel 1

Einleitung

1.1 Mobilitätsunterstützung als Grundlage für personenbezogene Dienste

In der aktuellen Entwicklung der Kommunikationstechnik ist ein Trend zu personenbezogenen Kommunikations- und Informationsdiensten zu beobachten. Einem Teilnehmer wird in Zukunft ein Spektrum von Diensten zur Verfügung gestellt werden, das dieser unabhängig von seinem aktuellen Aufenthaltsort und der dort verfügbaren Kommunikationsinfrastruktur nutzen können wird. Das heißt, daß ein Teilnehmer in für ihn gewohnter Weise und unabhängig vom benutzten Endgerät, von der Netzzugangstechnik und von den Vermittlungstechniken im Kernnetz auf Kommunikations- und Informationsdienste zugreifen können wird und für andere Teilnehmer entsprechend erreichbar sein wird. Dabei wird das Dienstespektrum und die Dienstgüte nur durch die Leistungsfähigkeit der verfügbaren Kommunikationsinfrastruktur begrenzt werden.

Ein wichtiger Schritt hin zu personenbezogenen Kommunikationsdiensten wurde mit der Entwicklung öffentlicher Mobilkommunikationsnetze getan. GSM, das „*Global System for Mobile Communications*“ (siehe Kapitel 2.3), sei hier als weltweit erfolgreichstes System dieser Art exemplarisch genannt. Mit Hilfe eines Teilnehmer-Identifikationsmoduls, das in der Regel in ein mobiles Endgerät eingeführt wird, gibt sich ein Teilnehmer einem Kommunikationsnetz zu erkennen. Ein Wechsel in das Netz eines anderen GSM-Netzbetreibers ist möglich („*Roaming*“). Nach dem Einbuchen kann der Teilnehmer auf Kommunikationsdienste zugreifen und ist über das mobile Endgerät erreichbar. Auch die Mobilität des Endgeräts wird im Rahmen des GSM-Standards unterstützt. Bisher ist für GSM-Teilnehmer nur eine Auswahl von Diensten aus einem standardisierten Dienstespektrum über Netzgrenzen hinweg zugänglich. Mit der Einführung von CAMEL (*Customized Applications for Mobile network Enhanced Logic*) wird es möglich, personenbezogene zusätzliche Dienstmerkmale über Netzgrenzen hinweg zur Verfügung zu stellen. Dies geschieht auf der Basis der Techniken des „Intelligenten Netzes“ (siehe Kapitel 2.4).

Dehnt man das Konzept der Teilnehmermobilität auf beliebige Endgeräte bzw. Netzzugangspunkte aus – ermöglicht man also zum Beispiel auch das Einbuchen an Endgeräten im Festnetz – so erhält man einen universellen personenbezogenen Telekommunikationsdienst. Solch ein Dienst wurde von der ITU-T als *Universal Personal Telecommunication* [F.850, F.851] stan-

standardisiert. Auch hierbei sollen personenbezogene Dienste und Dienstmerkmale netzeübergreifend angeboten werden und dabei vom Teilnehmer jederzeit beeinflusst werden können. Eine Fortführung dieses Konzepts findet sich in der Standardisierung zum sogenannten *Virtual Home Environment*, das zur Zeit im Rahmen des *Universal Mobile Telecommunication System* (UMTS, siehe Kapitel 2.3.3) entwickelt wird. Hier wird in besonderem Maße das sogenannte „*Look-and-Feel*“ des Dienstzugangs personenbezogen – und damit netze- und endgeräteübergreifend – konfigurierbar gemacht.

In einem nächsten Schritt werden Informationsdienste und typische Internet-Dienste, wie zum Beispiel *E-Mail*, Zugriff auf *World-Wide-Web*-Seiten und *Homebanking*, mit den klassischen Telekommunikationsdiensten zusammengeführt. Telefonieren am Rechner mit *Voice over IP*, das Laden von WWW-Seiten auf mobile Endgeräte sowie Bildtelefonie mittels eines ISDN-Endgerätes einerseits und einer Kamera am Rechner andererseits sind Beispiele für eine Konvergenz von Diensten. Für Telekommunikationsnetzbetreiber bietet sich hier die Chance, zusätzlich zur reinen Bereitstellung von Kommunikationsverbindungen die Authentisierung von Teilnehmern, die Entgelterfassung und die Rechnungsstellung für in Anspruch genommene Dienste durchführen zu können.

Personenbezogene Kommunikations- und Informationsdienste werden eine einheitliche Fassade bilden, hinter der die Heterogenität der Netze und der Netztechniken vor den Nutzern dieser Dienste verborgen wird. Damit können den Teilnehmern Dienste in der von ihnen gewünschten Weise zu jeder Zeit und an jedem Ort zur Verfügung gestellt werden. Die Realisierung dieser Dienste wird sich auf eine Reihe von Funktionen abstützen. Zum Teil werden diese Funktionen im Endgerät oder im Teilnehmer-Identifikationsmodul vorhanden sein, zum Teil werden sie von der Kommunikationsnetz-Infrastruktur erbracht werden, und zum Teil werden sie in Servern oder anderen speziellen Vorrichtungen außerhalb der eigentlichen Kommunikationsnetze bereitgestellt werden. Die verschiedenen Funktionen müssen zusammenwirken, damit eine Vielzahl personenbezogener Kommunikations- und Informationsdienste verfügbar gemacht werden kann.

Eine wesentliche Basis zur Erbringung dieser Dienste stellt die Unterstützung der Mobilität von Personen und Endgeräten dar. Eine effektive und effiziente Mobilitätsverwaltung muß dabei so angelegt werden, daß sie für eine wachsende Zahl von Teilnehmern in einer heterogenen Umgebung schnell und ressourcensparend arbeitet. Während die Mobilitätsverwaltung in Mobilkommunikationssystemen wie GSM für eine begrenzte Zahl von Teilnehmern, die sich im wesentlichen in einem einzigen Netz aufhalten, ausgelegt ist, wird Mobilitätsverwaltung in Zukunft der Tatsache Rechnung tragen müssen, daß ein sehr hoher Prozentsatz der Telekommunikationsteilnehmer mobil sein will und diese Mobilität sich über verschiedene Netze und Netztechniken erstrecken wird. Neben der Effizienz und der Skalierbarkeit sind aber auch Aspekte der Datensicherheit von wachsender Bedeutung. Vorbeugender Schutz vor Datenmißbrauch kann durch Dezentralisierung und Verteilung von Daten sowie vor allem durch Datensparsamkeit erzielt werden.

Eine Vielzahl neuer oder bestehender Netzzugangstechniken ermöglicht an verschiedenen Orten und auf unterschiedliche Weise Zugang zu Kommunikations- und Informationsdiensten. Dieser Zugang sollte personenbezogen möglich sein. Das Spektrum der Zugangstechniken reicht dabei von breitbandigem, drahtgebundenem Zugang mit Hilfe von ISDN/B-ISDN oder xDSL-Techniken (*Digital Subscriber Line*) über den (schnurlosen) Zugang zu lokalen Rechnernetzen (*Wireless LAN*) bis hin zu satellitengestützter Kommunikation. Eine wesentliche Rolle werden öffentliche Mobilkommunikationsnetze wie GSM und UMTS mit Piko-, Mikro- und Makro-Zellen spielen. Ein Teil der Endgeräte-Mobilität wird nach wie vor im Zugangssystem unterstützt werden. Die verschiedenen Ausprägungsformen des Ausrufens (*Paging*) von mobilen Endgeräten in zellularen Zugangssystemen macht diese erreichbar. Auch intelligente LAN-Brücken (*Bridge*) können Rechnermobilität in begrenztem Maße unterstützen.

Viele Veröffentlichungen befassen sich mit der Mobilität von Endgeräten innerhalb von Funkzellen oder zwischen Funkzellen, um daraus Rückschlüsse für den Entwurf von übertragungstechnischen Verfahren oder für die Mobilitätsverwaltung innerhalb von Zugangssystemen abzuleiten. Diese Arbeit befaßt sich mit dem Teil der Mobilität, der über die Begrenzungen eines einzelnen Zugangssystems hinausgeht, bei dem sich in der Regel die Adresse eines Endsystems in der Vermittlungsschicht ändert, mit Mobilität von Personen und Endgeräten, die zwischen Netzzugangspunkten und sogar zwischen Netzen stattfindet. Die wesentliche Aufgabe, die der Mobilitätsverwaltung hierbei zukommt, besteht darin, einerseits den mobilen Teilnehmer bzw. das mobile Endgerät für andere Teilnehmer erreichbar zu machen und andererseits dem mobilen Teilnehmer Zugriff auf die von ihm gewünschten Dienste zu ermöglichen. Anders als bei Mobilität innerhalb eines Zugangssystems spielt hier die verteilte Natur des betrachteten Systems eine wichtige Rolle. Dies gilt insbesondere für die verteilte Datenhaltung und die Verteilung von Ursprung und Ziel von Mobilitätsverwaltungsprozeduren.

Um netze- und netztechnikenübergreifende Mobilität für eine große Zahl von Teilnehmern zu ermöglichen, bedarf es leistungsfähiger und skalierbarer Verfahren zur Mobilitätsverwaltung. Bisherige Veröffentlichungen beschränken sich in der Regel auf einen Teilaspekt oder einen einzelnen Vorschlag, ohne einen Überblick über Gemeinsamkeiten und Unterschiede zwischen verschiedenen Netzen und Netztechniken zu geben. Die vorliegende Arbeit beschreibt daher zuerst, welche grundlegenden Mechanismen zur Unterstützung von Mobilität in Kommunikationsnetzen vorhanden sein müssen, und systematisiert dann die verschiedenen Optionen, die sich beim Entwurf der entsprechenden Verfahren bieten. Dies geschieht netztechnikenübergreifend und kann deshalb als eine wesentliche Grundlage für den Entwurf von Mobilitätsverwaltungskonzepten in zukünftigen heterogenen Netzstrukturen dienen. Die Tragfähigkeit der vorgestellten Systematisierung wird durch eine Reihe von neuen, leistungsfähigen Vorschlägen zur Mobilitätsverwaltung untermauert. Die Vorschläge sind dabei teilweise netztechnikspezifisch und teilweise netztechnikenübergreifend.

Die Leistungsfähigkeit neuer Vorschläge wird häufig durch Studien, die bestimmte Annahmen über das Teilnehmerverhalten machen, belegt. Ein direkter Vergleich zwischen verschiedenen neuen Vorschlägen ist in der Regel schwierig, da die zugehörigen Studien meist unterschiedliche Annahmen über das Teilnehmerverhalten machen und unterschiedliche Teilnehmermodelle zu Grunde legen. In dieser Arbeit wird ein Vorschlag für ein allgemein einsetzbares Teilnehmermodell gemacht, das alle wesentlichen Aspekte des Teilnehmerverhaltens berücksichtigt. Besonderer Wert wird dabei auf die Modellierung des Teilnehmerverhaltens in Bezug auf Bewegung und den Ursprung von Kommunikationsbeziehungen („kommende Rufe“) gelegt. Letzteres wird häufig nicht ausreichend beachtet. Darüberhinaus wurde das Modell so gestaltet, daß es relativ einfach parametrisiert werden kann und nicht übermäßig komplex ist, so daß es auch für große Systeme mit vertretbarem Aufwand verwendet werden kann. Es kann daher als Referenzmodell zum Vergleich verschiedener Mobilitätsverwaltungsverfahren eingesetzt werden.

Die Parametrisierung eines problemangepaßten Teilnehmermodells ist in der Regel eine komplexe Aufgabe, da sehr viele Parameter, die zum Teil in Abhängigkeit stehen, in geeigneter Weise gewählt werden müssen. In dieser Arbeit werden verschiedene, zum Teil sehr schnelle und leistungsfähige Verfahren zur Parametrisierung entwickelt und beschrieben. Besonders hervorzuheben sind hierbei die vorgeschlagenen Verfahren¹ zur Parametrisierung des Bewegungsmodells sowie die Einführung verschiedener Ruftypen bei der Modellierung der Teilnehmersuche, die den Ursprung von Rufen in Abhängigkeit vom Bewegungsverhalten der Teilnehmer beschreiben können.

Zusammen mit einem Lastmodell und einem flexiblen Netzmodell bildet die Teilnehmermodellierung eine wesentliche Voraussetzung für die Leistungsbewertung verschiedener Mobilitätsverwaltungskonzepte. Kapitel 6 illustriert diese Verwendungsmöglichkeit, indem exemplarisch einige Verfahren einer relativ einfachen Leistungsbewertung unterzogen werden. Dabei werden wichtige Erkenntnisse über den Einsatz dieser Verfahren für eine Fortentwicklung im Rahmen der GSM- oder der UMTS-Mobilitätsverwaltung gewonnen.

1.2 Übersicht über die Arbeit

Der Kern der Arbeit gliedert sich in zwei Teile. Der erste Teil besteht aus den Kapiteln 2 bis 4, der zweite Teil aus den Kapiteln 5 und 6. Beide Teile sind nach einem ähnlichen Schema aufgebaut. Zuerst werden jeweils ausführlich grundlegende Aspekte und Vorgehensweisen dargestellt, um dann jeweils am Ende die dabei gewonnenen Erkenntnisse und erarbeiteten Methoden exemplarisch anzuwenden. Der erste Teil der Arbeit definiert und systematisiert das Problemfeld Mobilität in Kommunikationsnetzen. Es werden netztechnische Grundlagen

1. Besonders elegant ist die Methode zur Bestimmung von Zustandsübergangswahrscheinlichkeiten bei vorgegebenen globalen Zustandsverlassensraten unter Verwendung von Gewichten (s. Kapitel 5.1.3.3 und 5.1.3.5).

beschrieben und verschiedene Möglichkeiten zur Unterstützung von Mobilität diskutiert. Als Abschluß wird eine Reihe von Vorschlägen zur Weiterentwicklung präsentiert. Der zweite Teil befaßt sich mit Methoden zur Modellierung von Teilnehmern, Verfahren und Netzen mit dem Ziel, durch eine Leistungsbewertung verschiedene Verfahren und Netzarchitekturen in quantitativer Form vergleichen zu können. Als Abschluß wird eine solche Leistungsbewertung exemplarisch durchgeführt, wobei Vorschläge aus dem ersten Teil der Arbeit aufgegriffen werden.

Kapitel 2 behandelt die netztechnischen Grundlagen. Zu Anfang wird das Phänomen *Mobilität* im Rahmen des OSI-Referenzmodells dargestellt sowie in Bezug zu Adressierung und Nummerierung im Internet und in klassischen Telekommunikationsnetzen gestellt. Danach werden wichtige unterstützende Netztechniken wie das Signalisiersystem Nummer 7 und das Konzept des „Intelligenten Netzes“ präsentiert. Daneben wird die Mobilitätsverwaltung in GSM, die zukünftige Entwicklung im Rahmen von UMTS und die Mobilitätsunterstützung in IP-basierten Netzen (*Mobile IP*) näher beschrieben.

Kapitel 3 befaßt sich mit Konzepten zur Mobilitätsverwaltung. Die Aufgaben der Mobilitätsverwaltung werden definiert und das Zusammenspiel von Wegführung und Zielsuche diskutiert. Der Schwerpunkt des Kapitels liegt auf der Klassifizierung verschiedener Verfahren zur Datenhaltung. Dabei werden die Aspekte *Plazierung von Daten*, *Ausfindigmachen von Daten* und *Datenaktualisierung* näher beleuchtet. Abschließend wird eine Reihe von bestehenden und vorgeschlagenen Verfahren zur Mobilitätsverwaltung beschrieben und eingeordnet.

Kapitel 4 präsentiert Vorschläge zur Weiterentwicklung bestehender und vorgeschlagener Verfahren. Die Vorschläge werden in drei Gruppen gegliedert: Vorschläge zur Unterstützung optimaler Nutzverkehrsführung, Vorschläge zur Fortentwicklung der GSM-Mobilitätsverwaltung und Vorschläge für eine netztechnikenübergreifende Mobilitätsverwaltung. Drei Vorschläge sind hervorzuheben: Das Senden von frühen *Binding Updates* in *Mobile IP*, die Verwendung von *Proxy HLRs* in GSM (bzw. UMTS) und die Einrichtung einer *Common Distributed Database* für netzeübergreifende Mobilitätsverwaltung.

In Kapitel 5 werden Methoden zur Modellierung des Teilnehmerverhaltens, der Verfahren zur Mobilitätsverwaltung (Lastmodell) und der verarbeitenden Instanzen (Netzmodell) vorgeschlagen. Das Teilnehmermodell gliedert sich in ein Bewegungs- und ein Rufmodell. Ziel der Teilnehmermodellierung ist es, im Zusammenspiel mit dem Lastmodell geeigneten Modellverkehr zu erzeugen, um damit eine Leistungsbewertung der verschiedenen Verfahren zur Mobilitätsverwaltung durchführen zu können.

In Kapitel 6 schließlich werden einfache Kriterien und Verfahren zur Leistungsbewertung verwendet, um auf dieser Basis exemplarisch Vergleiche verschiedener Netzstrukturen und Mobilitätsverwaltungsverfahren durchzuführen. Kriterien sind dabei vor allem die anfallende Netzlast in Form von Signalisierverkehr und in Form von in Prozessoren und Datenbankknoten zu bearbeitenden Aufgaben sowie die Antwortzeiten elementarer Mobilitätsverwaltungsprozeduren. Die Arbeit schließt mit einer Zusammenfassung und einem Ausblick.

Kapitel 2

Netztechnische Grundlagen

Mobilitätsverwaltung wurde häufig erst nachträglich in Kommunikationssysteme eingebracht. Dadurch stützt sie sich im allgemeinen auf bereits vorhandene Protokolle und Architekturen ab und wird dabei wesentlich von ihnen geprägt. In diesem Kapitel werden die entsprechenden netztechnischen Grundlagen – soweit relevant – eingeführt und in Bezug zu Mobilität gestellt. Es werden wichtige Begriffe, Konzepte und Protokolle vorgestellt, auf die dann in anderen Kapiteln zurückgegriffen wird. Hierbei werden meist zuerst deutsche Begriffe verwendet, die soweit möglich dem Begriffswerk des ITG-Fachausschusses 5.2 [ITG96a, ITG96b, ITG97] entnommen sind. Danach werden die in der englischsprachigen Standardisierung verwendeten Bezeichnungen und die zugehörigen Akronyme genannt. Letztere sind Legion, was bei einer Vielzahl unterschiedlicher Standardisierungsaktivitäten in mehreren verschiedenen Organisationen wohl leider nicht zu vermeiden ist. Ein Abkürzungsverzeichnis ab Seite vii bietet eine Hilfestellung.

Im ersten Unterkapitel wird die Adressierung und Numerierung im OSI-Referenzmodell, im Internet und in klassischen Telekommunikationsnetzen – soweit sie Mobilität betreffen – dargestellt. Ein Teil über Mobilität im OSI-Referenzmodell gibt dabei eine Definition von Portabilität und Mobilität von OSI-Instanzen sowie von Endsystemen und zeigt auf, welche grundlegenden Mechanismen in Netzen vorhanden sein müssen, damit Mobilität unterstützt werden kann.

Das Signalisiersystem Nummer 7 wird im folgenden Unterkapitel kurz eingeführt, da es die Grundlage für den Austausch von Steuerinformationen in GSM und im „Intelligenten Netz“ darstellt und da es Mechanismen zur netzeübergreifenden Adressierung von Instanzen der Mobilitätsverwaltung und der Dienstleistung bietet. GSM wird als Beispiel eines Mobilkommunikationssystems der zweiten Generation vorgestellt; UMTS als dessen Nachfolger der dritten Generation. Der Teilnehmerlokalisierung in GSM ist ein eigener Abschnitt gewidmet.

Die Standardisierung des sogenannten „Intelligenten Netzes“ definiert einen Rahmen zur Beschreibung von Architekturen zur verteilten Erbringung von netzbasierten Zusatzdiensten, wie etwa auch der Mobilitätsunterstützung. Aktuelle und zukünftige Netze werden wichtige Steuerfunktionen im Rahmen des „Intelligenten Netzes“ realisieren. Es ist damit Grundlage für neue, flexible Kommunikationsdienste wie etwa die personenbezogene Telekommunikation und wird deshalb im dritten Unterkapitel ausführlich dargestellt. Außerdem wird auch eine

kurze Beschreibung von TINA (*Telecommunications Information Networking Architecture*) gegeben, da TINA der weiteren Entwicklung des „Intelligenten Netzes“ wichtige Impulse bieten kann. Abschließend wird Mobilitätsunterstützung im Rahmen des „Intelligenten Netzes“ diskutiert.

Das vierte Unterkapitel stellt das Konzept von *Mobile IP* vor, das Mobilität im Internet ermöglichen soll. Verbindungslose Kommunikation in der Vermittlungsschicht, andere Konzepte zur Dienstbringung und größere Möglichkeiten zu unerwünschten Manipulationen unterscheiden das Internet unter anderem von den klassischen Telekommunikationsnetzen. Mobilitätsverwaltung im Internet ist bisher nur in einer sehr rudimentären Form spezifiziert, bietet jedoch neue Mechanismen und Fragestellungen sowie interessante Perspektiven zur weiteren Entwicklung.

2.1 Adressierung, Numerierung und Mobilität

Mobilitätsunterstützung ist zu einem großen Teil ein Adressierungsproblem. Im Folgenden soll die Adressierung in wichtigen Kommunikationssystemen kurz dargestellt und in Bezug zu Mobilität gestellt werden.

2.1.1 Adressierung und Mobilität bei OSI

Will im OSI-Referenzmodell eine allgemeine (N+1)-Instanz eine andere (N+1)-Instanz adressieren, so verwendet sie dazu eine (N)-Adresse, die einen oder mehrere gleichartige (N)-Dienstzugangspunkte (*Service Access Point, SAP*) adressiert [X.650] (siehe dazu auch Abbildung 2-1). Kennt die (N+1)-Instanz diese (N)-Adresse nicht, so greift sie auf ein (N+1)-Adreßbuch

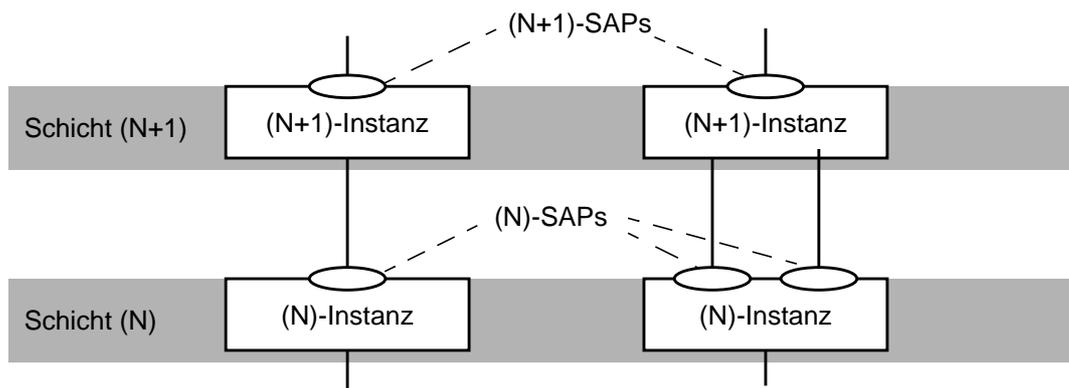


Abbildung 2-1: Zuordnung von (N)-Instanzen und (N)-Dienstzugangspunkten zur entsprechenden Schicht.

oder auf einen (N+1)-Verzeichnisdienst zurück, welche zum eindeutigen Namen (*title*) einer (N+1)-Instanz eine oder mehrere (N)-Adressen bereitstellen. Eine (N)-Dienstdateneinheit mit der (N)-Adresse wird dann an die darunterliegende Schicht übergeben, wo eine (N)-Instanz die (N)-Adresse wiederum auf den Namen einer (N)-Instanz oder direkt auf eine Adresse eines (N-1)-Dienstzugangspunktes abbildet. Häufig wird dabei eine (N-1)-Adresse direkt aus einer

(N)-Adresse gewonnen. Dies soll an einem Beispiel erläutert werden. Eine Anwendungsinstanz wird durch eine andere mit Hilfe einer P-Adresse (Adresse der Darstellungsschicht) angesprochen. Diese P-Adresse besteht zumeist aus einer N-Adresse (Adresse der Vermittlungsschicht), die um Selektoren für die Wahl der S-Instanz (T-Selektor, Selektor der Transportschicht), der P-Instanz (S-Selektor, Selektor der Sitzungsschicht) und der A-Instanz (P-Selektor, Selektor der Darstellungsschicht) erweitert ist. Folglich kann in den Schichten 6, 5 und 4 die jeweilige (N-1)-Adresse aus der jeweiligen (N)-Adresse gewonnen werden. In der Vermittlungsschicht dagegen wird zur N-Adresse die nächste zuständige N-Instanz ermittelt, welche dann über eine DL-Adresse (Adresse der Sicherungsschicht) angesprochen wird. Tabelle 2-1 zeigt die zu diesem Beispiel gehörigen Adressen und Abbildung 2-2 das zugehörige Adressierungskonzept.

P- Adresse:	P-Selektor, S-Selektor, T-Selektor, N-Adresse
S-Adresse:	S-Selektor, T-Selektor, N-Adresse
T-Adresse:	T-Selektor, N-Adresse
N-Adresse:	N-Adresse
DL-Adresse:	DL-Adresse

Tabelle 2-1: Beispiel für Adressen der Schichten 2 bis 6 des OSI-Referenzmodells

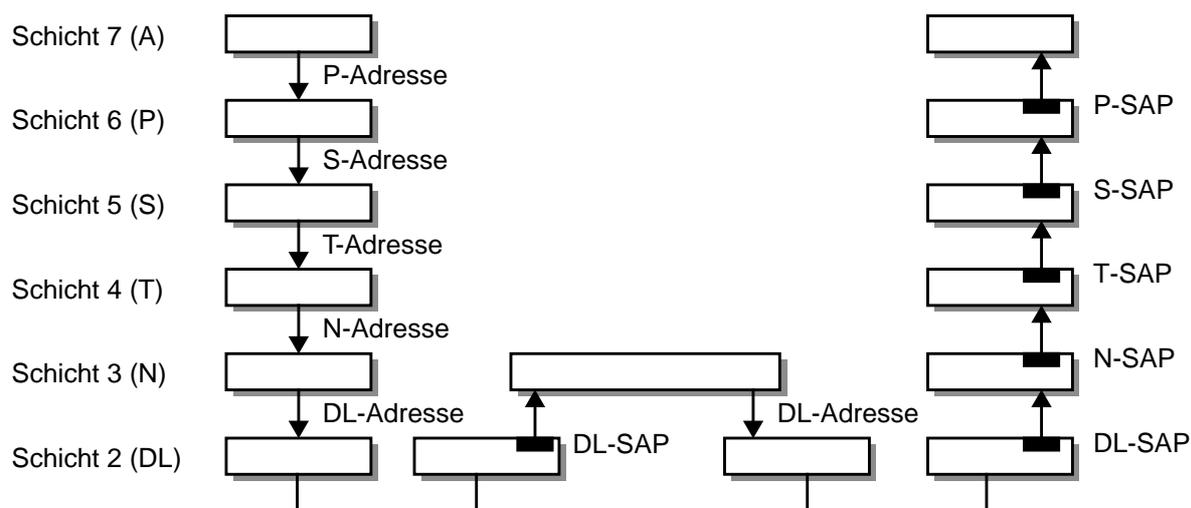


Abbildung 2-2: Adressierung einer Anwendungsinstanz durch eine andere in den Schichten 2 bis 6 des OSI-Referenzmodells.

Mobilität wird in den OSI-Spezifikationen nicht explizit behandelt. In dieser Arbeit wird deshalb dargestellt, in welcher Weise das Phänomen Mobilität mit Hilfe von OSI-Terminologie beschrieben werden kann. Dazu wird zwischen der Mobilität von (N)-Instanzen und der Mobilität von Endsystemen unterschieden.

Mobilität von Instanzen der Schicht (N) bedeutet, daß eine (N)-Instanz die (N-1)-Dienstzugangspunkte, unter denen sie erreichbar ist, ändern kann, ohne daß die (N)-Beziehungen (*associations*), die sie unterhält, unterbrochen werden. Werden (N)-Instanzen prinzipiell mit ihrem (ortsunabhängigen) Namen bezeichnet, so kann mit Hilfe einer entsprechend flexiblen Abbildung von (N)-Namen auf (N-1)-Adressen Mobilität in der Schicht (N) unterstützt werden.

Daneben müssen Mechanismen bestehen, um zu einer (N)-Beziehung gehörige Verbindungen bei einem Wechsel der Dienstzugangspunkte neu zu konfigurieren. Fehlen solche Mechanismen, so können nur Beziehungen, die auf verbindungslose Dienste aufbauen, aufrecht erhalten werden. Werden bei einem Zugangspunktwechsel Kommunikationsbeziehungen zwischen Anwendungsinstanzen nicht wiederhergestellt, so spricht man nicht von Mobilität, sondern von Portabilität.

Das obige Beispiel zur Adressierung zeigt, daß in den Schichten 4 bis 6 (N)-Instanzen meist direkt über eine (N-1)-Adresse angesprochen werden, die aus einer (N)-Adresse abgeleitet wird, was Mobilität von Instanzen dieser Schichten verhindert. In der Anwendungsschicht wird zur Ermittlung von Anwendungsinstanzadressen meist auf einen Verzeichnisdienst zugegriffen. Damit ist zumindest Portabilität von Instanzen in der Anwendungsschicht möglich.

Mobilität einer Instanz der Vermittlungsschicht würde nach obiger Definition heißen, daß sie ihren DL-Dienstzugangspunkt ändert. Dies ist zwar prinzipiell möglich, das Problem von Mobilität und Adressierung tritt auf den Schichten 1 bis 3 jedoch anders in Erscheinung. Zumeist werden ja nicht (N)-Instanzen, sondern ganze Endsysteme bewegt, deren (N)-Instanzen ihre Adressen beibehalten sollen. Was sich unter Umständen verändert, ist die Konnektivität der (N)-Instanzen, das heißt die Partner der (N)-Dienstprotokolle (*Adjacent Layer Protocols*) ändern sich.

Ändert sich das Übertragungsmedium, so werden bestehende Schicht-2-Beziehungen unterbrochen und die Partnerinstanzen der Vermittlungsschicht können nicht mehr direkt über ihre bisherigen DL-Adressen erreicht werden. Hier werden nun Protokolle benötigt, um die Adressen möglicher Partnerinstanzen der Vermittlungsschicht zu ermitteln und mit diesen über neue Schicht-2-Beziehungen Kontakt aufzunehmen¹. Ziel muß es sein, daß das Endsystem unter seinen bisherigen N-Adressen wieder erreichbar wird. Haben sich die Partnerinstanzen der Vermittlungsschicht nicht geändert, so kann eine lokale Aktualisierung von Verkehrslenkungstabellen erfolgen. Bei flacher Adressierung wäre es auch denkbar, die Verkehrslenkungstabellen global mit einem flexiblen und leistungsfähigen Verkehrslenkungsprotokoll entsprechend zu modifizieren. Aus Effizienzgründen verwenden größere Netze jedoch im allgemeinen hierarchische Adressierung und Verkehrslenkung, bei der nur ein Teil der Adresse ausgewertet wird. Will man dies beibehalten, so muß die Erreichbarkeit des Endsystems auf andere Weise wiederhergestellt werden. Ist das Endsystem, und damit auch dessen Instanzen, wieder erreichbar, so müssen vorher bestehende Verbindungen wiederaufgenommen werden. Dies gilt insbesondere für Verbindungen der Vermittlungsschicht, die aus einer Verkettung von Beziehungen zwischen Vermittlungsinstanzen bestehen. Diese Verkettung muß rekonfiguriert werden. Im einfachsten Fall wird von der alten N-Partnerinstanz eine Weiterführung zum neuen Ort veranlaßt.

1. Im Idealfall werden die zugehörigen Schicht-2-Verbindungen parallel unterhalten und nach dem Wechsel der Verkehrsführung werden die nicht mehr benutzten Schicht-2-Verbindungen abgebaut. Um die Darstellung einfach zu halten, wird hier jedoch von einem Abreißen der Verbindung ausgegangen.

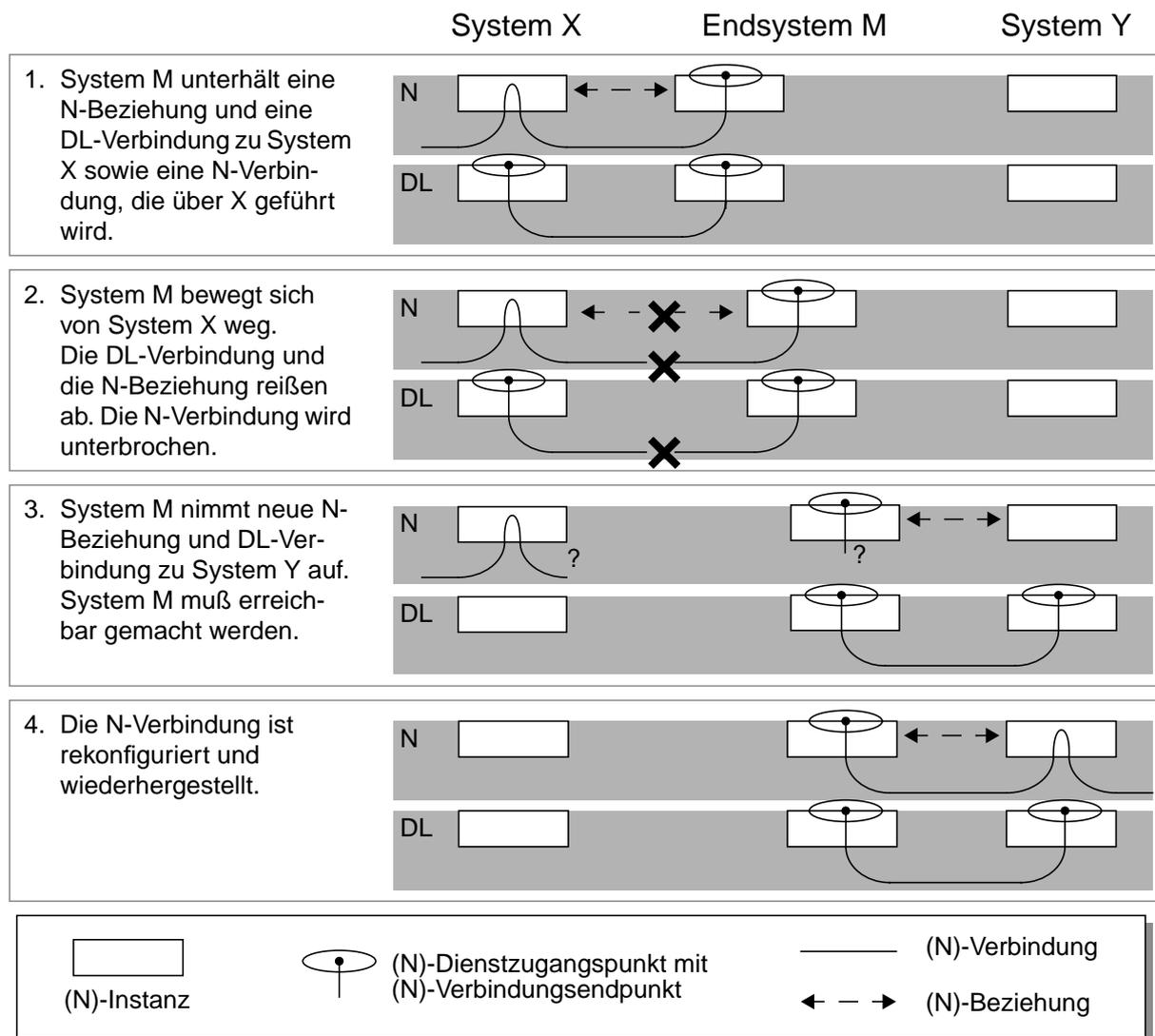


Abbildung 2-3: Denkbare Ablauf für Mobilitätsunterstützung bei OSI, wenn ein Endsystem M bei bestehender N-Verbindung das physikalische Medium wechselt und das zur N-Verbindung gehörige nächste Transit- oder Endsystem sich ändert.

Abbildung 2-3 veranschaulicht den denkbaren Ablauf für Mobilitätsunterstützung bei OSI. Endsystem M wechselt das physikalische Medium bei bestehender N-Verbindung über das Transit- oder Endsystem X. Dabei reißen die DL-Verbindung und die N-Beziehung zu X ab. Die N-Verbindung wird zeitweise unterbrochen. Zu System Y baut M eine neue DL-Verbindung und eine neue N-Beziehung auf, während die N-Verbindung unterbrochen bleibt. Schließlich wird die N-Verbindung neu konfiguriert und damit die N-Verbindung über Y zu M geführt.

Bei hierarchischer Adressierung und entsprechender Verkehrslenkung ist es nicht möglich, Verkehrslenkungstabellen endsystemspezifisch zu modifizieren. Hier bietet sich die Technik zweifacher Adreßumsetzung an [BPT96]. Ein Endsystem, das nicht mehr unmittelbar unter seiner ursprünglichen N-Adresse erreichbar ist, nimmt am neuen Ort eine temporäre N-Adresse an. Wird in der Vermittlungsschicht erkannt, daß es sich um ein mobiles Endsystem handelt, so wird die permanente N-Adresse auf die temporäre umgesetzt und diese zur Verkehrslenkung verwendet. Sobald klar ist, daß keine weitere Verkehrslenkung erforderlich ist,

wird die temporäre N-Adresse wieder auf die permanente N-Adresse umgesetzt und damit der entsprechende N-Dienstzugangspunkt adressiert. Damit bleibt die Adreßumsetzung den höheren Schichten verborgen. Wo die jeweiligen Umsetzungen stattfinden, ist abhängig vom verwendeten Protokoll zur Unterstützung von Endsystemmobilität.

Hier zeigt sich das fundamentale Problem von Mobilität in bestehenden Systemen. Zur Adressierung von Endsystemen bzw. N-Dienstzugangspunkten werden Adressen benutzt, die in der Vermittlungsschicht zur Verkehrslenkung verwendet werden. Erst die Trennung von N-Dienstzugangspunktadresse und Verkehrslenkungsinformation ermöglicht Mobilität. In den meisten heutigen Systemen, die Mobilität unterstützen, wird eine N-Dienstzugangspunktadresse bis zu dem Punkt im Netz zur Verkehrslenkung verwendet, an dem erkannt wird, daß es sich um ein mobiles Endgerät handelt. Dort findet dann eine Adreßumsetzung statt, die es ermöglicht, mit Hilfe der Verkehrslenkung zum mobilen Ziel zu gelangen.

Auch auf den Schichten 1 und 2 kann Endsystemmobilität in begrenztem Maße ermöglicht werden. Insbesondere in lokalen Netzen mit Vielfachzugriff auf ein gemeinsames physikalisches Medium ist begrenzte Mobilität von Endsystemen möglich. Unter Beibehaltung des genutzten Mediums bleibt ein Endsystem über seine MAC-Adresse erreichbar. Die Verwendung von Signalregeneratoren (*repeater*) oder Brücken (*bridge*) erweitert den Erreichbarkeitsbereich in beschränktem Maße, wobei die Filtermechanismen in intelligenten Brücken den Standortwechsel durch Beobachtung des Verkehrs bemerken. In dedizierten, funkbasierten Mobilkommunikationssystemen wird die Mobilität von Endgeräten auch durch Leistungssteuerung (*power control*) und intelligente Antennen, die ihre Richtcharakteristik an den aktuellen Aufenthaltsort anpassen, unterstützt. Daneben wird in solchen Systemen die Weitergabe von Schicht-2-Verbindungen durch spezielle *Handover*-Verfahren (inklusive der Ausnutzung von Makrodiversität) und mit Hilfe spezieller Zugangsnetze ermöglicht.

2.1.2 Adressierung und Mobilität im Internet

Für Adressierung und Mobilität gilt im Internet Ähnliches wie beim OSI-Referenzmodell. IP-Adressen sind Adressen der Vermittlungsschicht. Adressen der Transportschicht sind IP-Adressen, die mit einer Protokollkennung erweitert werden. Eine Internet-Anwendungsadresse ist eine IP-Adresse plus Protokollkennung plus Port-Nummer. Tabelle 2-2 zeigt dies beispielhaft.

Anwender verwenden häufig Bezeichner wie Rechnernamen oder E-Mail-Adressen zur Adressierung. Solch ein Bezeichner muß dann von einer Internet-Anwendung auf eine IP-Adresse abgebildet werden. Dazu bedient sich diese im allgemeinen der Dienste des *Domain Name System* (DNS). Anfragen an das DNS bestehen aus einem *domain name* und aus einer Beschreibung des Typs des gesuchten Datums.

Anwendungsadresse:	Port-Nummer, Protokollkennung, IP-Adresse
T-Adresse:	Protokollkennung, IP-Adresse
N-Adresse:	IP-Adresse
DL-Adresse:	MAC-Adresse

Tabelle 2-2: Beispiel für Adressen im Internet

Ein *domain name* wird einem Namensraum entnommen, der Baumstruktur hat. Jedem Knoten des Baumes ist ein Bezeichner zugeordnet, der aus einer Folge von maximal 63 Zeichen besteht.² Der Wurzelknoten hat einen Bezeichner der Länge Null. Die Bezeichner der direkten Kinder des Wurzelknotens sind standardisiert. Beispiele sind com, de, org und edu. Ein vollständiger *domain name* ist eine Folge von Bezeichnern der Baumstruktur vom bezeichneten Knoten bis zum Wurzelknoten, die jeweils durch einen Punkt voneinander getrennt sind. `www.ind.uni-stuttgart.de`. beispielsweise bezeichnet einen einzelnen Rechner. Anfragen an das DNS können auch relativ erfolgen. Der *domain name* wird dann nicht bis zum Wurzelbezeichner angegeben (z. B.: `www.ind.uni-stuttgart` oder `www.ind.uni-stuttgart.de`). Im einfachsten Fall wird die IP-Adresse unter Angabe des Bezeichners für einen einzelnen Rechner erfragt. Häufig wird jedoch mit dem Bezeichner auch ein Dienst in einer bestimmten *domain* angesprochen. Beispielsweise wird eine Anfrage nach Rechnern, die für den E-Mail-Dienst einer *domain* (z.B.: `ind.uni-stuttgart.de`.) zuständig sind, vom DNS mit einer Liste von Rechnernamen beantwortet, für welche dann wiederum die IP-Adressen erfragt werden können.

Hier könnte sich eine gewisse Portabilität von Rechnern erreichen lassen, indem die Zuordnung vom Rechnernamen auf die IP-Adresse im DNS verändert wird. Das DNS basiert auf einer verteilten Datenhaltung, die für häufige Zugriffe und seltene Änderungen optimiert ist [BPT96]. So ist die Gültigkeitsdauer einer Datensatzkopie (Cache-Eintrag) meist auf 1 bis 7 Tage eingestellt. Änderungen werden erst nach Ablauf der Gültigkeitsdauer durch Neuansforderung bekannt. Die heutige Realisierung dieses Systems ist für große Mobilität deshalb nicht geeignet. Ein Mechanismus, der Änderungen global bekannt gibt, ist bisher nicht vorgesehen, da er unter Umständen zu einem zu hohen Verkehrsaufkommen führen würde.

Die Unterstützung von Rechnermobilität unter Beibehaltung aller Adressen ist im Internet mit ähnlichen Problemen konfrontiert wie Endsystemmobilität im OSI-Referenzmodell. Da jedoch die Vermittlungsschicht verbindungslos arbeitet, ist eine Rekonfigurierung der Verbindungsführung hier nicht notwendig. Das Problem, bei hierarchischer Adressierung und zugehöriger Verkehrslenkung mobile Rechner auf der Vermittlungsschicht erreichbar zu halten, bleibt jedoch bestehen. *Mobile IP*, ein möglicher Lösungsansatz für dieses Problem, wird in Kapitel 2.5 mit seinen Vor- und Nachteilen näher beschrieben.

2. Zulässige Zeichen sind die 26 Buchstaben des Alphabets, die Ziffern 0 bis 9 und der Bindestrich. Groß- und Kleinschreibung wird nicht unterschieden. Ziffern dürfen nicht am Anfang stehen. Ein Bindestrich darf weder am Anfang noch am Ende stehen.

2.1.3 Numerierung und Adressierung in klassischen Telekommunikationsnetzen

ISDN und GSM haben sich aus dem analogen Telefonnetz entwickelt, wo Nummern³ zur Adressierung von festen Teilnehmeranschlüssen verwendet werden. Telekommunikationsnummern unterliegen einem weltweit eindeutigen Numerierungsplan (E.164, [E.164]) und dürfen maximal 15 relevante Ziffern haben. Verkehrsausscheidungsziffern, die auf Grund der verwendeten offenen Numerierung notwendig sind, sowie Verbindungsnetzbetreiberkennzahlen zur verbindungsbezogenen Auswahl eines Netzbetreibers werden nicht zur relevanten Rufnummer gerechnet, beschränken jedoch den Numerierungsraum. Die Numerierung unterstützt hierarchische Verkehrslenkung. Die häufigsten Nummern sind geographische Nummern. Eine solche Nummer ist unterteilt in die Landeskenzahl, die nationale Zielkenzahl in Form einer Ortskenzahl und die Teilnehmerrufnummer (siehe Abbildung 2-4).

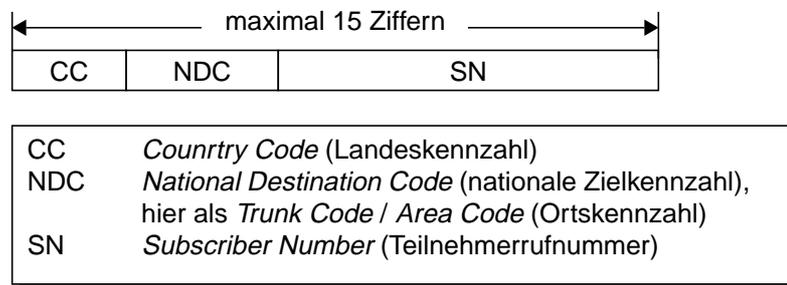


Abbildung 2-4: Struktur einer geographischen Nummer nach [E.164].

Nummern werden inzwischen jedoch nicht mehr ausschließlich zur Adressierung von festen Teilnehmeranschlüssen, sondern zunehmend auch zur Adressierung von Diensten, mobilen Endgeräten und mobilen Teilnehmern verwendet. So kann die Landeskenzahl oder die nationale Zielkennzahl als Dienstzugangskennzahl verwendet sein. Für Mobilfunknetze wird häufig die nationale Zielkennzahl als Netzzugangskennzahl benutzt. Darüberhinaus geben Nummern in der Regel für den Benutzer einen Hinweis auf die anfallenden Kosten eines Rufes.

Telekommunikationsnummern sind vergleichbar mit Adressen der Vermittlungsschicht im OSI-Referenzmodell. Auch sie werden zur Verkehrsführung verwendet. Mobilität und neue Dienste haben Nummern jedoch inzwischen auch zu Bezeichnern gemacht, die erst in zur Verkehrsführung geeignete Nummern übersetzt werden müssen. Dabei hat sich eine Situation ergeben, bei der solch ein Bezeichner nicht unbedingt als ein solcher erkannt wird, sondern zuerst zur Verkehrsführung verwendet wird, bevor eine Nummernumsetzung stattfindet. Als Beispiel sei der Anruf zu einem Mobilfunkteilnehmer genannt. Dabei wird der Ruf aus einem fremden Netz anhand der Landeskenzahl und/oder der Netzzugangskennzahl bis ins Heimatnetz des Teilnehmers geführt und dort in eine international gültige (temporäre) Rufnummer seines Endgerätes (Aufenthaltsrufnummer) umgesetzt. Ähnliches gilt für Dienste des „Intelligenten Netzes“, bei dem die Dienstkennung dazu verwendet wird, den Ruf zu einer Dienstzugangsvermittlungsstelle zu führen, wo dann eine Rufnummernumsetzung erfolgen kann.

3. Folge von Symbolen aus [0-9].

Ein zweiter Trend beeinträchtigt die Tauglichkeit von Rufnummern zur Verkehrsführung. Gemäß Telekommunikationsgesetz §42(5) [Wit96] haben Betreiber von Telekommunikationsnetzen „sicherzustellen, daß Nutzer bei Wechsel eines Betreibers und Verbleiben am selben Standort ihnen zugeteilte Nummern beibehalten können (Netzbetreiberportabilität)“. Damit werden also auch beim Verbleiben eines Teilnehmeranschlusses am selben Ort Mechanismen notwendig, um, ähnlich wie bei Mobilität, Informationen zur Führung einer Verbindung, die nicht unmittelbar aus der gewählten Nummer abgeleitet werden können, zu erlangen.

Um ineffiziente Verbindungsführung zu vermeiden, ist es sinnvoll, möglichst früh zu erkennen, daß eine gewählte Nummer nicht zur Verkehrsführung geeignet ist. Auf internationaler Ebene werden dazu bestimmte Landeskennzahlen für dienst- oder personenbezogene Nummern reserviert. So sollen beispielsweise die Landeskennzahl 800 für *International Freephone* und die Landeskennzahlen 878 und 879 für internationale persönliche Rufnummern verwendet werden. Auf nationaler Ebene [SB97] werden bestimmte Netzzugangskennzahlen als Dienstkennzahlen verwendet, so zum Beispiel in Deutschland 700 für persönliche Rufnummern sowie 800 und 801 für entgeltfreies Anrufen (bisher: 130). In Frankreich und Spanien werden dazu in der Regel die Numerierungsgassen 6 bis 9 von geographischen Nummern freigehalten. Großbritannien beschränkt geographische Nummern sogar auf die Numerierungsgasse 1. In der GSM-Phase 2+ wird ein Dienstmerkmal vorbereitet, bei dem GSM-Netze erkennen können, ob ein Ruf zu einem Mobilfunkteilnehmer eines fremden Netzes geführt werden soll, so daß bereits im Ursprungsnetz eine entsprechende Datenbankabfrage mit anschließender Rufnummernübersetzung stattfinden kann. Auf der Ebene eines einzelnen Netzes muß es auch zu einer Aufteilung des Numerierungsraumes kommen. So muß ein Mobilfunknetzbetreiber erkennen können, ob die gerufene Nummer bei einem von außen kommenden Ruf bereits die Aufenthaltsrufnummer eines Teilnehmers ist, oder ob sie erst noch umgesetzt werden muß. Um den Rufnummernbedarf für Aufenthaltsrufnummern klein zu halten, wird solch eine Nummer meist für jeden kommenden Rufaufbau neu zugeteilt. Wird sie dagegen für die gesamte Aufenthaltsdauer an einem Ort vergeben, so belegt jeder eingebuchte Teilnehmer dauerhaft zwei Nummern, nämlich seine permanente Rufnummer und eine Aufenthaltsrufnummer.

Mittelfristig zeichnet sich die persönliche Telekommunikationsnummer ab, die unabhängig vom benutzten Endgerät, vom Dienstanbieter, vom Netzbetreiber und vom aktuellen Wohnort einem Teilnehmer zugeordnet ist. Dabei kann ein Teilnehmer unter Umständen auch an mehreren Endgeräten registriert sein und in Abhängigkeit von Dienstanforderungen, Preis oder anderen Kriterien über eines oder mehrere dieser Endgeräte angesprochen werden.

2.2 Das Signalisiersystem Nummer 7

Signalisierung dient zum Austausch von Steuerungsinformationen. Die Signalisierung zwischen Netzknoten in modernen Telekommunikationsnetzen wird durch das Signalisiersystem Nummer 7 [Q.700] unterstützt, welches es erlaubt, Nachrichten unabhängig von Nutzkanalver-

bindungen über ein paketorientiertes Datennetz – das Signalisierernetz – auszutauschen. Hohe Zuverlässigkeit, Robustheit und Echtzeitfähigkeit des Signalisierernetzes werden durch entsprechende Protokollmechanismen sichergestellt. In Abbildung 2-5 wird die Protokollarchitektur des Signalisiersystems Nummer 7 im ISDN und im B-ISDN gezeigt und in Beziehung zu den sieben Schichten des OSI-Referenzmodells gestellt. Beispiele weiterführender Literatur sind [KPS94, LTU95, BTE88, Baf95, Zepf95, Man91].

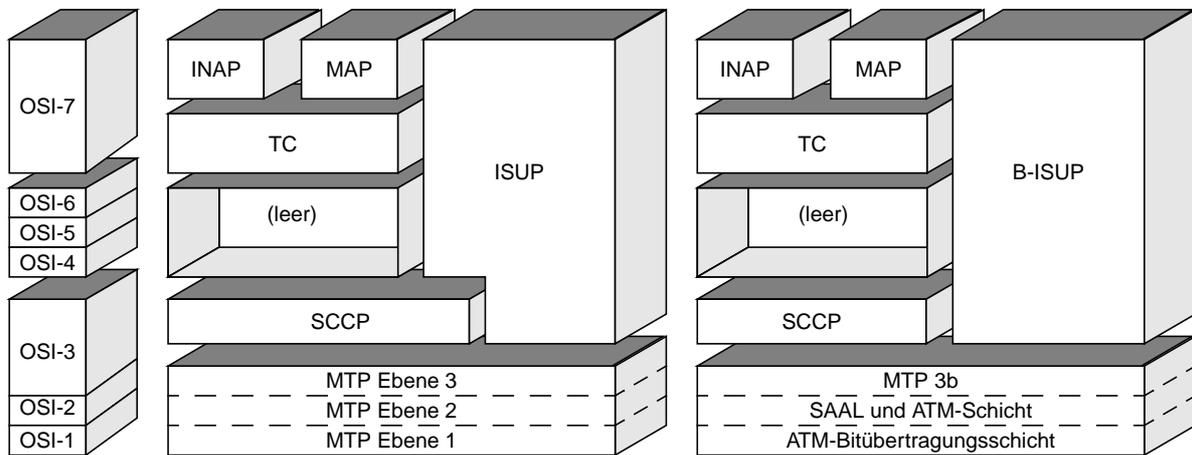


Abbildung 2-5: Protokollarchitektur des Signalisiersystems Nummer 7 im ISDN und im B-ISDN in Relation zu den sieben Schichten des OSI-Referenzmodells.

Im ISDN unterstützt das Signalisiersystem Nummer 7 mit Hilfe des ISDN-Anwenderteils (*ISDN User Parts*, ISUP) [Q.761] im wesentlichen den abschnittswiseen Auf- und Abbau von Nutzkanalverbindungen sowie den Ende-zu-Ende Austausch von Steuernachrichten zwischen Ursprungs- und Zielvermittlungsstelle zur Realisierung zusätzlicher ISDN-Dienstmerkmale. Im B-ISDN ist hierfür der B-ISUP vorgesehen [Q.2761]. Neben dieser hauptsächlich nutzkanalbezogenen Signalisierung gewinnt die nutzkanalunabhängige Signalisierung zwischen Vermittlungsstellen, Netzdatenbanken und Dienststeuerknoten im Rahmen neuer, verteilter Anwendungen zunehmend an Bedeutung. So bildet das Signalisiersystem Nummer 7 zum Beispiel die Grundlage für die Realisierung der Mobilitätsverwaltung in zellularen Mobilkommunikationssystemen wie GSM (*Mobile Application Part*, MAP) sowie für die Diensterbringung im Rahmen des „Intelligenten Netzes“ (*Intelligent Network Application Part*, INAP). In beiden Beispielen erfolgt – unabhängig von der Nutzkanalführung – ein Austausch von Steuerinformationen sowohl zwischen Vermittlungsstellen und Datenbank- bzw. Dienststeuerknoten als auch zwischen den verschiedenen Datenbank- bzw. Dienststeuerknoten.

Das Signalisiersystem Nummer 7 unterstützt solche verteilten Anwendungen durch die anwendungsunabhängigen *Transaction Capabilities* (TC) [Q.771], welche einerseits die Etablierung eines Dialoges und das Aushandeln eines Anwendungskontextes erlauben als auch andererseits innerhalb eines solchen Dialoges den Austausch von Anwendungsprotokolldateneinheiten (APDUs) zum Aufruf entfernter Operationen und zum Zurückliefern ihrer Ergebnisse ermöglichen. Die *Transaction Capabilities* sind in der Anwendungsschicht des OSI-Referenzmodells anzusiedeln und weisen in weiten Teilen Übereinstimmung mit den OSI-Anwen-

dungsdienstelementen (*Application Service Element, ASE*) ACSE (*Association Control Service Element, X.217, [X.217]*) und ROSE (*Remote Operation Service Element, X.219, [X.219]*) auf.

Im Gegensatz zu diesen OSI-Protokollen stützen sich die *Transaction Capabilities* des Signalisiersystems Nummer 7 nicht auf verbindungsorientierte Dienste ab, sondern auf die verbindungslosen Dienste⁴ des sogenannten *Signalling Connection Control Parts* (SCCP) [Q.711], welche den verbindungslosen Diensten der OSI Schicht 3 entsprechen. Dies wird den typischen Anwendungen aus den Bereichen Mobilkommunikation und Dienststeuerung gerecht, da dort im allgemeinen nur geringe Mengen an Steuerdaten mit möglichst kleiner Verzögerung übermittelt werden müssen.

Der *Signalling Connection Control Part* gestattet es, Benutzer der *Transaction Capabilities* weltweit zu adressieren. Dies geschieht auf der Basis von Subsystemnummern (SSN) und sogenannter *Global Titles* – allgemeinen Adressen, die nicht unmittelbar zur Adressierung innerhalb des Signalisiernetzes verwendet werden können. Der SCCP setzt diese Global Title an gegebener Stelle in verwertbare Adressen um (*Global Title Translation, GTT*). Zur Erbringung der SCCP-Dienste stützt dieser sich auf den Nachrichtentransferteil (*Message Transfer Part, MTP, [Q.701]*) ab. Der MTP ist lediglich in der Lage, Knoten innerhalb des eigenen Netzes, also typischerweise innerhalb des Netzes eines Netzbetreibers, zu adressieren. Netzeübergreifender Nachrichtenaustausch geschieht via SCCP-Instanzen, die an Netzübergängen platziert sind und Verkehrslenkung und Adreßumsetzung anhand der *Global Title* durchführen.

In GSM beispielsweise wird zur Adressierung der Heimatdatei (*Home Location Register, HLR*) eines Teilnehmers bei einer Aufenthaltsortsaktualisierung die Subsystemnummer 00000110_b (zur Adressierung des HLR-ASEs) und ein *Global Title* gemäß E.214 [E.214] verwendet, der aus der Internationalen Mobilteilnehmerkennung (*International Mobile Subscriber Identity*⁵, IMSI) gemäß E.212 [E.212] durch Umsetzung in Landeskennzahl (*Country Code*) und Netzzugangskennzahl (*Network Code*) gemäß E.164 [E.164] und unter Beibehaltung der Mobilteilnehmerkennung (*Mobile Subscriber Identification Number, MSIN*) gebildet wird. Hat die Nachricht ihren Ursprung in einem fremden Netz, so wird sie via mehrere SCCP-Instanzen anhand von *Country Code* und *Network Code* ins Heimatnetz des Teilnehmers geführt. Dort wird anhand von MSIN und SSN das für den Teilnehmer zuständige HLR ermittelt und die Nachricht via MTP zur dortigen SCCP-Instanz übermittelt.

4. Für den SCCP sind auch verbindungsorientierte Dienste spezifiziert, die jedoch im allgemeinen nicht implementiert sind.

5. In den noch gültigen ITU-T Empfehlungen der E-Serie von 1988 wird statt *Subscriber* der Begriff *Station* gebraucht. In der GSM-Standardisierung bei ETSI hat sich jedoch der Begriff *Subscriber* durchgesetzt, da ja nicht ein spezifisches Endgerät (*Station*) sondern ein Teilnehmer (*Subscriber*) bezeichnet werden soll. Der Teilnehmer wiederum ist in GSM über das Teilnehmer-Identifikationsmodul (*Subscriber Identity Module*), welche in verschiedenen Endgeräten verwendet werden kann und die IMSI beinhaltet, definiert.

Der MTP bietet einen verbindungslosen Nachrichtenübermittlungsdienst an, bei dem Nachrichten zum selben Ziel, die eine gemeinsame Kennung (*Signalling Link Selection*) tragen, den selben Weg durch das Signalisiernetz nehmen und aufgrund von abschnittsweiser Reihenfolgesicherung ihre Reihenfolge beibehalten.⁶ Daneben beinhaltet er wichtige Mechanismen zur Aufrechterhaltung des Signalisiernetzbetriebs bei Überlast und bei Verfügbarkeitsveränderungen (Ausfall bzw. Wiederinbetriebnahme von Knoten und Signalisierstrecken). Im klassischen Signalisiersystem Nummer 7 besteht der *Message Transfer Part* aus drei Ebenen. Die Ebene 3 hat Verkehrslenkungs- und Netzmanagementaufgaben. Ebene 2 führt Reihenfolgesicherung über Signalisierstrecken durch, die in der Ebene 1 typischerweise eine Übertragungsrate von 64 kbit/s haben. Im Breitband-ISDN wird die Funktionalität der MTP Ebene 3 beibehalten und an die darunterliegenden Protokolle angepaßt [CS95]. Die Signalisierstrecken werden hier durch ATM-Verbindungen gebildet, bei denen die SAAL (*Signalling ATM Adaptation Layer*) im wesentlichen die Aufgaben der MTP Ebene 2 übernimmt.

2.3 GSM und UMTS

2.3.1 GSM – Global System for Mobile Communications

GSM ist ein öffentliches Mobilkommunikationssystem der zweiten Generation, das sich vom europäischen Standard quasi zu einem Weltstandard entwickelt hat. Ende 1999 waren laut GSM MoU Association über 300 GSM-Netze in 142 Ländern mit mehr als 215 Millionen Teilnehmern in Betrieb.⁷ Ende 1997 gab es in Deutschland 7,8 Millionen Mobilfunkteilnehmer in Netzen mit zellularer Technik. Dies sind 101 je 1000 Einwohner [ITS98]. Finnland dagegen hatte 419 Mobilfunkteilnehmer je 1000 Einwohner, Norwegen 383 und Schweden 358 [ITS98]. Selbst in den Vereinigten Staaten, wo eine Vielzahl etablierter analoger und digitaler Systeme GSM Konkurrenz machen, betrug Ende 1997 die Zahl der GSM-Teilnehmer 1,4 Millionen [ITS98].⁸

GSM ist ein digitales, zellulares System, das Sprach- und Datendienste bei netzeübergreifender Erreichbarkeit und automatischer Übergabe von Verbindungen bei Zellwechsel bietet. Die Mehrheit der Netze (z.B. die D-Netze in Deutschland) operieren im Frequenzbereich um

6. Bei Ausfall von Knoten oder Signalisierstrecken kann es in Ausnahmefällen zu Reihenfolgeveränderungen, Meldungsverlusten und/oder Duplizierungen kommen.

7. Quelle: GSM MoU Association, <http://www.gsmworld.com>, GSM World Newsletter, Issue 4, January 1998.

8. Insgesamt gab es dort jedoch zu diesem Zeitpunkt bereits 55 Millionen Mobilfunkteilnehmer (206 je 1000 Einwohner)

900 MHz (*GSM900*). Darüberhinaus werden (z.B. beim deutschen E-Plus-Netz) die Bereiche um 1800 MHz (*GSM1800*)⁹ und in den Vereinigten Staaten um 1900 MHz (*GSM1900*)¹⁰ für Netze verwendet, die sich kaum von den GSM900-Netzen unterscheiden.¹¹

Das Akronym GSM steht heute für *Global System for Mobile Communications*¹². Ursprünglich stand es für *Groupe Spécial Mobile*, die Bezeichnung für eine 1982 gegründete Arbeitsgruppe der CEPT, aus der die gleichnamige Gruppe beim 1988 gegründeten Europäischen Institut für Fernmeldenormen (ETSI, *European Telecommunications Standards Institute*) hervorging, die den GSM-Standard schließlich erarbeitet hat. Die GSM-Standardisierung erfolgt in Phasen. In Phase 1, die Ende 1990 abgeschlossen wurde, wurden Übermittlungsdienste für Sprache (13 kbit/s) und Daten (9,6 kbit/s) spezifiziert. Neben den darauf aufbauenden Telediensten, wie Telefondienst und Telefax, wurden in den Phasen 1 und 2 der Kurznachrichtendienst (*Short Message Service*, SMS) und eine Reihe von ISDN-ähnlichen zusätzlichen Dienstmerkmalen spezifiziert. Die Phase 2 wurde 1995 abgeschlossen und stellt im wesentlichen eine Überarbeitung der Phase 1 unter Hinzufügung neuer Zusatzdienste dar. In der aktuellen sogenannten Phase 2+ der Standardisierung werden eine Reihe von neuen Themen und Diensten diskutiert, deren Behandlung oft unabhängig voneinander erfolgt, und die auch unabhängig voneinander verabschiedet und eingeführt werden. Dazu gehören die Verbesserung der Sprachqualität durch verbesserte Sprachkodierung (*Enhanced Full Rate Codec* und *Adaptive Multi-Rate Codec*), ein neuer paketvermittelter Datendienst mit auf Paketebene konkurrierendem Zugriffsverfahren und einer eigenen Infrastruktur im Festnetz (*General Packet Radio Service*, GPRS) sowie ein höherbitratiger, durchschaltvermittelter Datendienst (*High Speed Circuit Switched Data*, HSCSD). Daneben werden weitere zusätzliche Dienstmerkmale spezifiziert und Konzepte des „Intelligenten Netzes“ eingebracht (*Customized Applications for Mobile network Enhanced Logic*, CAMEL). Insbesondere für Eisenbahnbetreiber¹³ werden Übertragungsverfahren und Protokolle für Geschwindigkeiten bis 500 km/h sowie Leistungsmerkmale des Bündelfunks¹⁴ entwickelt (GSM-R). Vertiefende Grundlagen zu GSM, insbesondere zu den Phasen 1 und 2, finden sich in [MP92, EV97].

9. auch: Digital Cellular System DCS 1800.

10. auch: Personal Communications System, PCS.

11. Durch den unterschiedlichen Frequenzbereich ergeben sich Veränderungen an Endgeräten und Sendeeinrichtungen. Die Mehrzahl der Endgeräte arbeitet derzeit nur in einem der drei Frequenzbereiche und kann daher in Netzen mit abweichenden Frequenzen nicht verwendet werden.

12. Abweichend davon, jedoch weniger gebräuchlich: *Global System for Mobile Communication* [G01.04].

13. in Zusammenarbeit mit der UIC, *Union Internationale des Chemins de Fer*.

14. GSM tritt damit in Konkurrenz zu digitalen Bündelfunksystemen wie z.B. TETRA (Trans European Trunked Radio System). Es wird dabei in der GSM-Phase 2+ nachträglich Funktionalität hinzugefügt, die bei diesen Systemen bereits in der Konzeption berücksichtigt wurde.

[G01.02] teilt die Netzarchitektur eines GSM-Netzes in die folgenden drei sogenannten Subsysteme auf:

- ❑ das Funk-Zugangssystem (*Base Station System, BSS*)
- ❑ das Vermittlungs- und Mobilitätsverwaltungs-Subsystem (*Switching and Management Subsystem, SMSS*)
- ❑ und das Betriebs- und Wartungs-Subsystem (*Operation and Maintenance Subsystem, OMSS*)

Abbildung 2-6 stellt die prinzipielle Architektur eines GSM-Netzes mit den wesentlichen Komponenten des BSS und des SMSS dar.

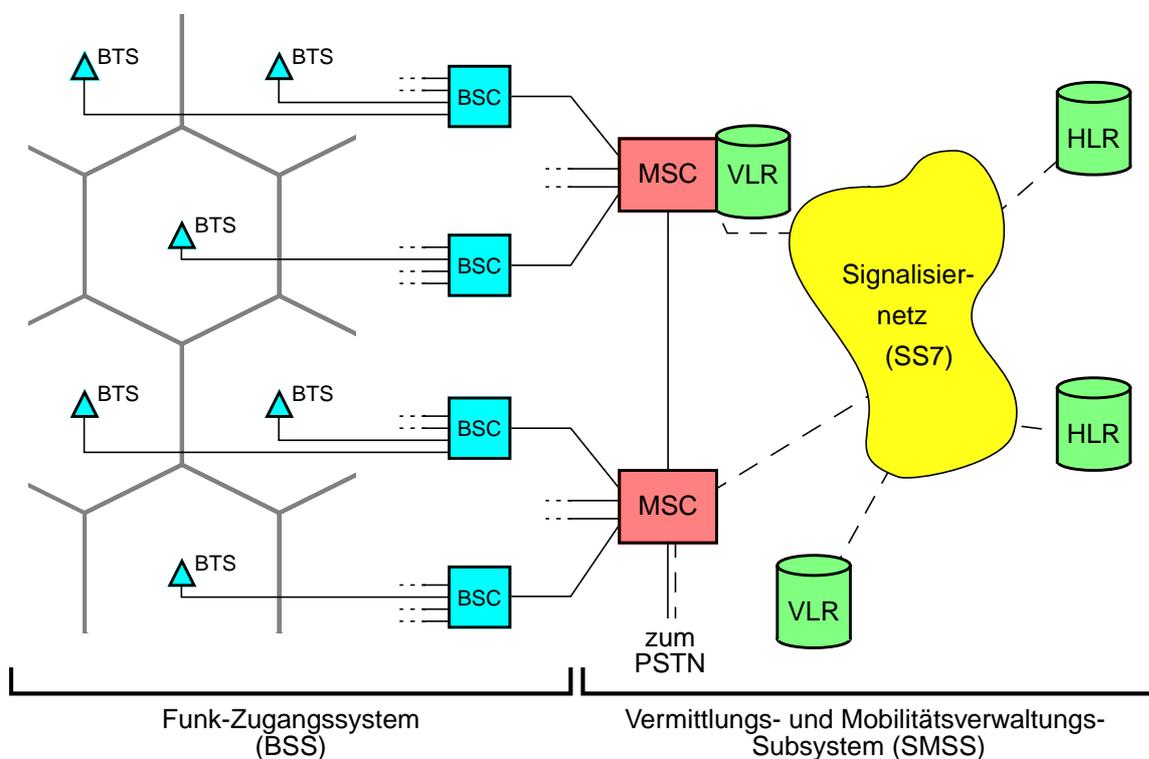


Abbildung 2-6: Prinzipielle Architektur eines GSM-Netzes.

Das Funk-Zugangssystem ermöglicht mobilen Endgeräten den Zugang zu einer Mobilfunk-Vermittlungsstelle (*Mobile Switching Centre, MSC*). Damit stellt es im wesentlichen ein Zugangsnetz dar. Es ist zuständig für die Funkkanalverwaltung und für die Anpassung der speziellen GSM-Luftschnittstelle an die ISDN-verwandte¹⁵ Schnittstelle zur MSC. Mehrere ortsfeste Sende- und Empfangseinrichtungen (*Base Transceiver Stations, BTS*), deren Versorgungsgebiete man als Funkzellen bezeichnet, werden von einer Basisstationssteuerung (*Base Station Controller, BSC*), mit der sie über Kabel oder Richtfunk verbunden sind, koordi-

15. Innerhalb des BSS und zur MSC hin finden eine Reihe von (zum Teil leicht abgewandelten) ISDN-Protokollen sowohl des UNI als auch des NNI Anwendung. Dies betrifft sowohl die Nutzkanäle als auch die Signalisierung. Zwischen BTS und BSC beispielsweise wird LAPD verwendet; zwischen MSC und BSC ein vereinfachter MTP und der SCCP.

niert. Die Steuerung von Funkkanalwechsellern für bestehende Verbindungen (*Handover*) innerhalb derselben Funkzelle oder zwischen Funkzellen, die sich unterhalb derselben BSC befinden, kann im BSS erfolgen.

Das Vermittlungs- und Mobilitätsverwaltungs-Subsystem stellt im Grunde ein erweitertes ISDN-Festnetz dar. Dies sieht man besonders deutlich daran, daß im BSS ein GSM-Sprachverkehrskanal mit typischerweise 13 kbit/s auf einen ISDN-B-Kanal mit 64 kbit/s umgesetzt wird. Auch der Auf- und Abbau von Rufen zwischen Mobilfunk-Vermittlungsstellen erfolgt wie im ISDN auf der Basis des ISUP. Neue Instanzen in diesem Netz sind die Besucherdateien (*Visitor Location Register, VLR*) und die Heimatdateien (*Home Location Register, HLR*), die untereinander und mit den MSCs eine verteilte Mobilitätsverwaltung auf der Basis des *Mobile Application Part* (MAP) [G09.02] realisieren. Der MAP stützt sich, wie oben beschrieben, auf die *Transaction Capabilities* und den SCCP des Signalisiersystems Nummer 7 ab. In der Nomenklatur unterscheiden die GSM-Standards nicht so präzise zwischen funktionalen und physikalischen Instanzen, wie das in der Standardisierung zum „Intelligenten Netz“ geschieht. Betrachtet man die Mobilitätsverwaltung in GSM auf der verteilten funktionalen Ebene des konzeptionellen Modells des IN¹⁶, so stellen MSC, HLR, VLR und das mobile Endgerät (*Mobile Station, MS*) funktionale Instanzen dar, zwischen denen Informationsfluß stattfindet. In [G01.02] werden die Schnittstellen zwischen diesen Instanzen mit den Großbuchstaben am Anfang des Alphabets bezeichnet. Abbildung 2-7 zeigt die entsprechenden Zusammenhänge.

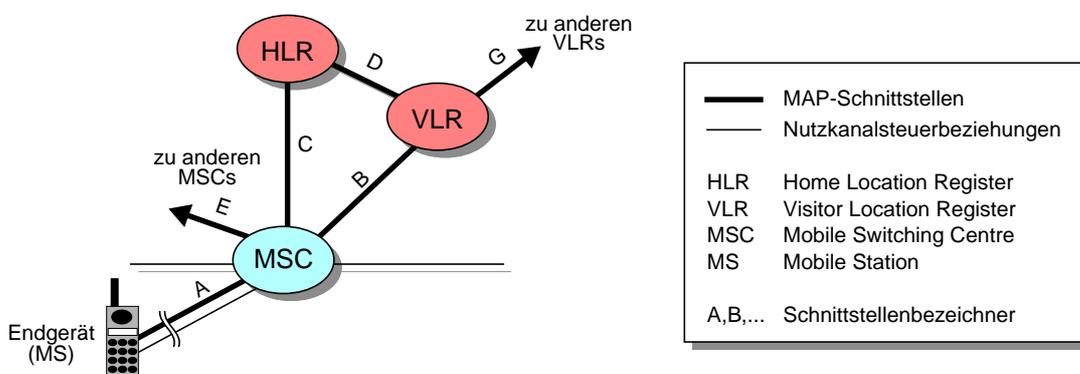


Abbildung 2-7: Funktionale Instanzen der Mobilitätsverwaltung und zugehörige Schnittstellen.

Sowohl die Schnittstellenbezeichner als auch die Bezeichnungen für die funktionalen Instanzen werden in der GSM-Standardisierung auch in der physikalischen Ebene für physikalische Instanzen und Schnittstellen verwendet. Häufig befinden sich VLR- und MSC-Funktionalität in einer Mobilfunk-Vermittlungsstelle, aber auch andere Zuordnungen zwischen funktionalen Instanzen und Netzknoten sind denkbar.¹⁷

16. Siehe dazu Kapitel 2.5.1.

17. So zum Beispiel HLR und VLR oder HLR, VLR und MSC in einem Netzknoten.

Das Betriebs- und Wartungs-Subsystem enthält typische Funktionen zum Netzmanagement. Daneben unterstützt es das Sicherheitsmanagement in GSM, wozu vor allem die Benutzer-authentifizierung gehört. In einem Berechtigungszentrum (*Authentication Centre*) wird für jeden Teilnehmer ein geheimer Schlüssel gehalten, der sich auch auf dem Teilnehmer-Identifikationsmodul (*Subscriber Identity Module*, SIM) des Teilnehmers befindet. Bei Bedarf wird im Berechtigungszentrum eine Liste mit mehreren Authentifizierungsparameter-Sätzen für ein *Challenge-Response*-Verfahren erzeugt, die im Netz von VLR zu VLR weitergereicht werden kann. Da jeder Parametersatz nur einmal benutzt werden darf, muß eine neue Liste angefordert werden, sobald alle Sätze verbraucht sind. Daneben ist im OMSS eine GeräteKennungsdatei (*Equipment Identity Register*) vorgesehen, mit deren Hilfe es beispielsweise ermöglicht werden soll, daß gestohlene Endgeräte deaktiviert werden.

2.3.2 Teilnehmerlokalisierung in GSM

Die Verfahren, die es ermöglichen, einen GSM-Teilnehmer ausfindig zu machen, werden im folgenden als *Teilnehmerlokalisierung* bezeichnet. Die Teilnehmerlokalisierung in GSM erfolgt in verteilter, hierarchischer Weise. Zwischen den in Abbildung 2-7 dargestellten funktionalen Instanzen findet dazu Informationsfluß statt. Eine Gruppe von Funkzellen wird zu einem Aufenthaltsbereich (*location area*) zusammengefaßt. Eine Besucherdatei (VLR) ist für einen oder mehrere Aufenthaltsbereiche zuständig und beinhaltet zu jedem Zeitpunkt die Information, welche Teilnehmer sich in welchen Aufenthaltsbereichen befinden. In der Heimatdatei (HLR) eines Teilnehmers ist verzeichnet, welche Besucherdatei gerade für den Teilnehmer zuständig ist. In jeder Funkzelle wird die Kennung des Aufenthaltsbereichs (*Location Area Identity*, LAI) ausgestrahlt, so daß ein Endgerät beim Wechsel der Funkzelle auch den möglichen Wechsel des Aufenthaltsbereichs feststellen kann. Erkennt ein Endgerät solch einen Wechsel, so initiiert es eine Ortsdatenaktualisierung (*Location Update*), indem es an die für den neuen Aufenthaltsbereich zuständige Besucherdatei eine entsprechende Meldung schickt. Handelt es sich dabei um eine andere Besucherdatei als jene, welche für den zuvor besuchten Aufenthaltsbereich zuständig war, so wird die Heimatdatei über den Wechsel der zuständigen Besucherdatei informiert.

Soll ein Ruf zu einem mobilen Teilnehmer aufgebaut werden, so wird an der ersten Vermittlungsstelle, die erkennt, daß es sich um einen mobile Teilnehmer handelt, die Aufenthaltsrufnummer des Endgeräts (*Mobile Station Roaming Number*, MSRN) benötigt, mit deren Hilfe ein weiterer Rufaufbau auf Basis des ISUP – oder anderer Zwischenamtssignalisierung – erfolgen kann. Dazu wird die Heimatdatei befragt, die dazu wiederum eine auf Rufbasis vergebene Aufenthaltsrufnummer von der zuständigen Besucherdatei anfordert. Dort wird die MSRN auf Rufbasis vergeben, um damit die Größe des Adreßraumes zur Adressierung von Teilnehmern an einer MSC klein zu halten. Erreicht ein Ruf daraufhin die Mobilfunk-Vermittlungsstelle, über die der Teilnehmer erreichbar ist, so wird er in allen Funkzellen seines Aufenthaltsbe-

reichs ausgerufen (*Paging*). Reagiert sein Endgerät auf dieses Ausrufen, so kann die zuständige Funkzelle identifiziert werden, und es erfolgt weitere Signalisierung zum Endgerät bezüglich eines kommenden Rufes.

Abbildung 2-8 zeigt vereinfacht den Protokollablauf für einen Aufenthaltsortwechsel, bei dem die zuständige Besucherdatei sich ändert. Ein Endgerät (MS) schickt eine Meldung zur Ortsdatenaktualisierung an die neu zuständige Besucherdatei (VLR_{neu}). Diese Meldung wird unverschlüsselt übertragen. Um die Identität des Teilnehmers zu verschleiern, wird deshalb eine von der vorherigen Besucherdatei (VLR_{alt}) vergebene, temporäre Identifikationsnummer (*Temporary Mobile Subscriber Identity*, TMSI) verwendet. Anhand der Kennung des vorherigen Aufenthaltsbereichs (LAI_{alt}) wird die vorherige Besucherdatei ermittelt und um die tatsächliche Identität des Teilnehmers (*International Mobile Subscriber Identity*, IMSI) sowie um Authentifizierungsparametersätze gebeten. Wird der Teilnehmer daraufhin erfolgreich authentifiziert, so erfolgt die Aktualisierung der Heimatdatei (HLR) unter Angabe der IMSI sowie der Adressen der neuen MSC und des neuen VLRs. Dabei wird auch der Teilnehmerdatensatz an die Besucherdatei übermittelt. Daneben wird der Eintrag in der vorherigen Besucherdatei gelöscht. Nach dem Aushandeln von Parametern zur Verschlüsselung und der Vergabe einer neuen TMSI wird die Ortsdatenaktualisierung dem Endgerät bestätigt.

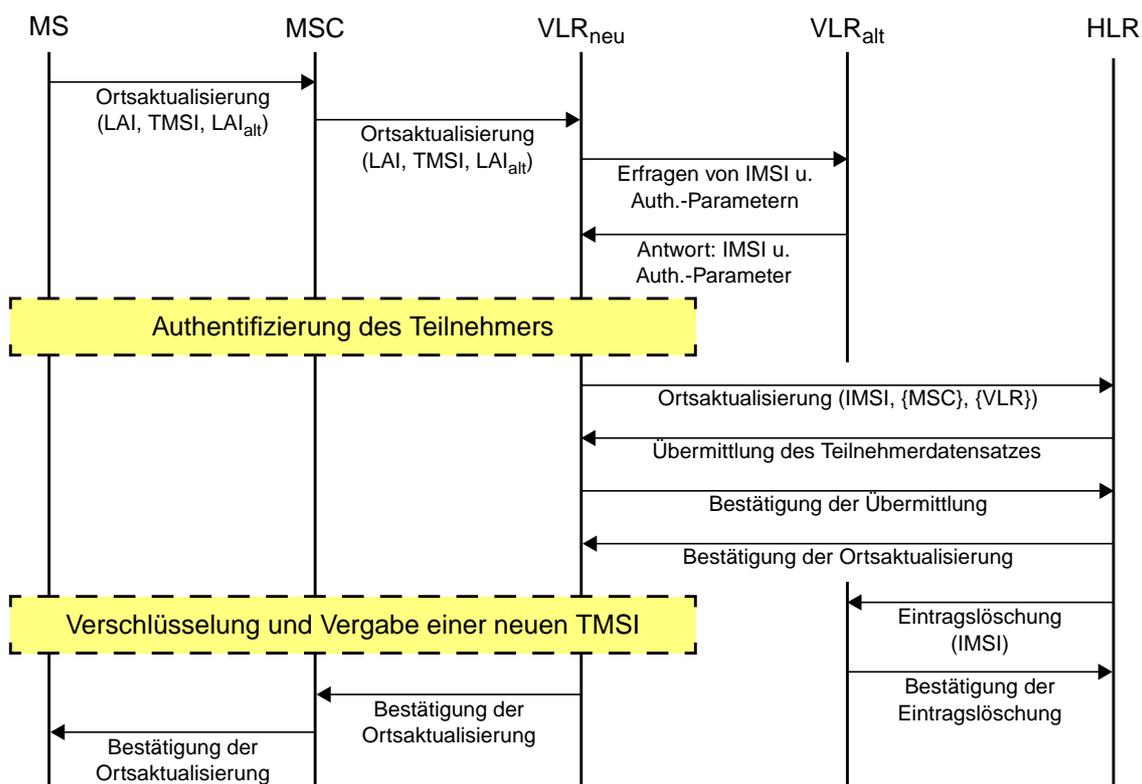


Abbildung 2-8: Protokollablauf beim Wechsel der zuständigen Besucherdatei (VLR) unter Verwendung der temporären Teilnehmererkennung (TMSI).

Abbildung 2-9 zeigt den Aufbau eines Rufes zu einem mobilen Teilnehmer. Adressiert wird dieser mit seiner öffentlichen Teilnehmerrufnummer, der *Mobile Subscriber ISDN Number*¹⁸ (MSISDN). Auf Basis dieser Rufnummer wird begonnen, den Ruf wie einen gewöhnlichen PSTN/ISDN-Ruf aufzubauen, bis erkannt wird, daß es sich um einen Ruf zu einem mobilen

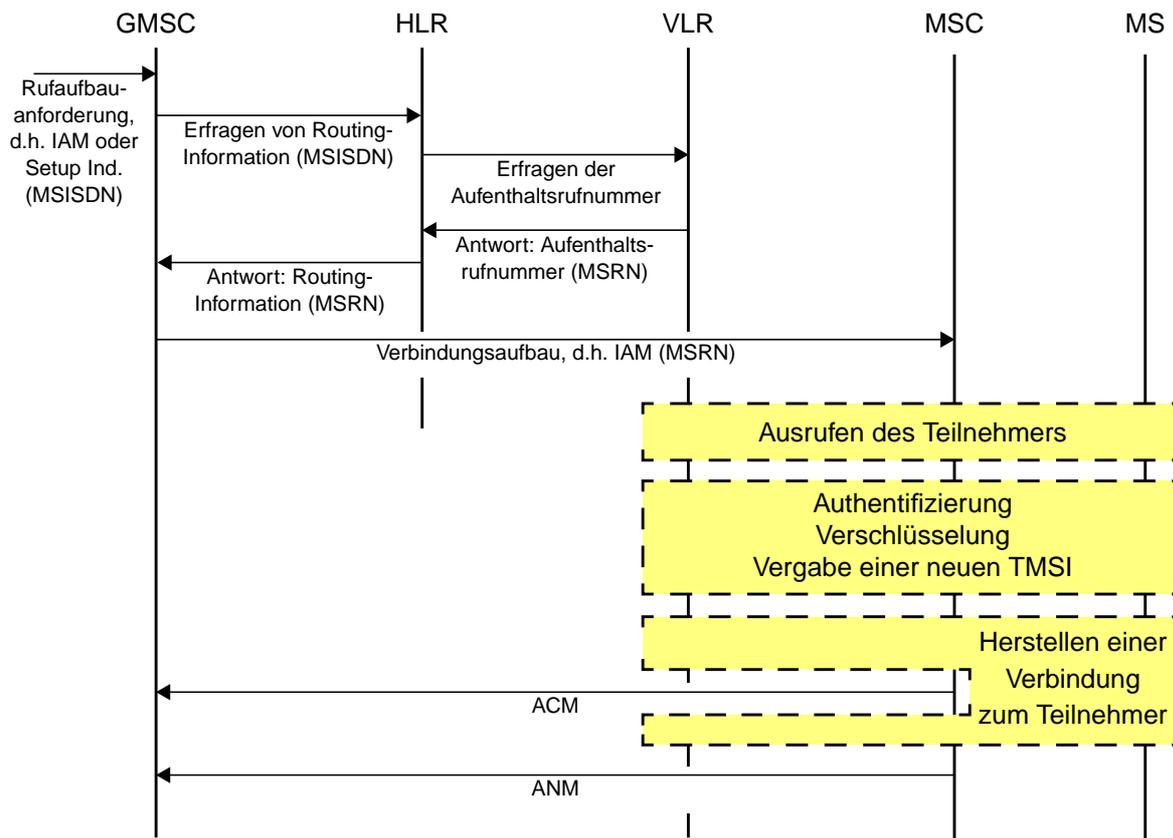


Abbildung 2-9: Protokollablauf beim Aufbau eines Rufes zu einem mobilen Teilnehmer.

Teilnehmer handelt. Liegt der Ursprung eines Rufes im Heimatnetz des gerufenen Teilnehmers, so erkennt dies zumeist die Ursprungs-Mobilfunk-Vermittlungsstelle. Ansonsten wird bisher ein Ruf in das Heimatnetz des Teilnehmers zu einer Vermittlungsstelle am Netzübergang (*Gateway MSC*, GMSC) geführt. Dort werden der Rufaufbauprozess angehalten und von der Heimatdatei des gerufenen Teilnehmers weitere Routing-Information angefordert. In der GSM-Phase 2+ soll das Erfragen dieser Information von fremden GSM-Netzen aus ermöglicht werden (*Support of Optimal Routing*, [G02.79, G03.79]). Ist der Teilnehmer erreichbar, so antwortet die Heimatdatei mit der aktuellen Aufenthaltsrufnummer (MSRN), die sie zuvor von der momentan zuständigen Besucherdatei anfordert. Mit Hilfe der MSRN wird der Ruf daraufhin von der Gateway MSC zur aktuell besuchten MSC aufgebaut. Dies geschieht durch gewöhnliche ISUP-Signalisierung, d.h. unter Verwendung einer *Initial Address Message* (IAM), auf die schließlich mit einer *Address Complete Message* (ACM) und einer *Answer Message* (ANM) geantwortet wird. Anhand der MSRN eines kommenden Rufes kann die MSC/VLR erkennen, welcher Teilnehmer adressiert wird und das Ausrufen und Authentifizieren des Teilnehmers vornehmen.

Die Mobilitätsverwaltung von GPRS unterscheidet sich deutlich von der Mobilitätsverwaltung für klassisches GSM. Sie ähnelt sehr stark der Mobilitätsverwaltung bei *Mobile IP*. Da die Konzepte von *Mobile IP* hier ausführlich dargestellt werden, wird auf eine genauere Beschrei-

18. GSM 01.04 [G01.04] definiert MSISDN als *Mobile Station ISDN Number*. Da jedoch der Teilnehmer adressiert wird und nicht ein spezielles Endgerät, wird hier der ebenfalls gängige Begriff *Mobile Subscriber ISDN Number* verwendet.

bung der Mobilitätsverwaltung von GPRS verzichtet. Wesentlich ist, daß für GPRS eine eigene, unabhängige Mobilitätsverwaltung durchgeführt wird, die auch das Endgerät dazu zwingt, Netzzugang und Mobilitätsverwaltung für die beiden unabhängigen Festnetzinfrastrukturen (klassisches GSM-Netz und GPRS- bzw. IP-Netz) getrennt zu organisieren. Idealerweise sollte solch eine doppelte Mobilitätsverwaltung in zukünftigen Systemen vermieden werden.

2.3.3 UMTS – Universal Mobile Telecommunication System

Als Nachfolgesystem von GSM wird bei ETSI zur Zeit das *Universal Mobile Telecommunication System* (UMTS) standardisiert. Entsprechende weltweite Aktivitäten werden unter dem Überbegriff IMT-2000 (*International Mobile Telecommunications 2000*¹⁹) zusammengefaßt. Als ein Mobilkommunikationssystem der dritten Generation soll es vor allem höhere Bandbreite für multimediale Telekommunikations- und Datendienste bereitstellen. Dazu wird zur Zeit insbesondere die Luftschnittstelle im Bereich von 2 GHz und das Zugangsnetz spezifiziert. Ziel ist es, bei großer geographischer Abdeckung zumindest eine dem ISDN-Basisanschluß entsprechende Bandbreite von 144 kbit/s oder sogar von 384 kbit/s pro Endgerät zur Verfügung stellen zu können. Eine Bandbreite bis zu 2 Mbit/s pro Endgerät in begrenzten Gebieten bei geringerer Geschwindigkeit soll ebenfalls unterstützt werden. Darüber, welche Dienste über die Luftschnittstelle abgewickelt werden sollen, herrscht zur Zeit noch eine gewisse Unklarheit. Wahrscheinlich werden ISDN-Dienste wie (qualitativ höherwertige) Sprach- und Bildtelefonie dazu gehören. Ebenso ist der Zugang zu zukünftigen Diensten eines Breitband-ISDN denkbar. Daneben werden sicherlich der Zugang zu Internet-Diensten sowie neue UMTS-spezifische Dienste zunehmend an Bedeutung gewinnen. Bei der Standardisierung von UMTS-spezifischen Festnetzaspekten wurde lange Zeit der Ansatz verfolgt, die Mobilitätsverwaltung für UMTS auf der Basis des „Intelligenten Netzes“ neu zu definieren. 1996 wurden die Standardisierungsaktivitäten bei ETSI jedoch neu strukturiert. Im *Global Multimedia Mobility Report* (GMM) der ETSI [GMM, BDW97] wird ein Standardisierungsrahmen vorgegeben, der eine Netzarchitektur mit folgenden vier konzeptionellen Bereichen definiert:

- ❑ Endgeräte (*terminals*),
- ❑ Zugangsnetze (*access networks*),
- ❑ Kerntransportnetze (*core transport networks*) und
- ❑ Anwendungsdienste (*application services*).

UMTS werden dabei vorwiegend die Bereiche Endgeräte und Zugangsnetze zugewiesen. Die Standardisierung für das Kernnetz, wozu auch die Dienststeuerung – die „Intelligenz“ – einschließlich der Mobilitätsverwaltung gehört, soll damit stärker im Rahmen von GMM und weniger UMTS-spezifisch erfolgen und sich aus ISDN beziehungsweise GSM entwickeln

19. Alte Bezeichnung: FPLMTS (*Future Public Land Mobile Telecommunication System*)

[U30.01]. Daraus resultiert nun auch das Bestreben, zur Mobilitätsverwaltung eine Weiterentwicklung der vorhandenen GSM bzw. GPRS-Protokolle einzusetzen [Dav97]. Ob dabei eine doppelte Mobilitätsverwaltung wie bei GSM/GPRS durchgeführt wird oder ob eine gemeinsame Mobilitätsverwaltung definiert werden wird, ist in der Standardisierung noch nicht entschieden. Da jedoch im Bereich der Kerntransportnetze bereits eine konzeptionelle Aufspaltung in durchschaltvermittelndes und IP-basiertes Netz vorgenommen wurde, ist zu befürchten, daß die doppelte Mobilitätsverwaltung à la GSM/GPRS auch bei UMTS zum Einsatz kommen wird, was eine Trennung der Mobilitätsverwaltung – und damit doppelten Aufwand – bis hin zu den Endgeräten zur Folge haben kann.

2.4 Das Konzept des „Intelligenten Netzes“

Telekommunikationsnetze sind im allgemeinen typische verteilte Systeme. Eigenschaften wie dezentrale Organisation und Autonomie von Teilsystemen, aber auch Heterogenität und evolutionäre Entwicklung zeichnen sie aus. Im ISDN beispielsweise werden Vermittlungsstellen verschiedener Hersteller zu einem Netz verknüpft. Basisdienste und Zusatzdienste²⁰ werden ausschließlich von den an einem Ruf beteiligten Vermittlungsstellen erbracht. Es werden dazu keine zentralen Instanzen benötigt. Will man in derselben Weise neue Dienste erbringen oder bestehende Dienstlogik und Dienstdaten verändern, so bedeutet dies, daß Dienstlogik und Dienstdaten in alle Vermittlungsstellen eingebracht bzw. verändert werden müssen, an denen die zugehörigen Dienstmerkmale verfügbar sein sollen. Wenn dies schnell, sicher und flexibel möglich sein soll, werden extrem hohe Anforderungen sowohl an die Software-Architektur der Vermittlungsstellen als auch an das Netzmanagementsystem gestellt. Aktuelle technische Entwicklungen wie die *Java Virtual Machine* [LY96] und TMN (*Telecommunication Management Network*, [M.3000]) lassen erkennen, daß eine derartige Architektur zumindest prinzipiell denkbar wäre. Anfang der achtziger Jahre, als das Konzept des „Intelligenten Netzes“²¹ (IN, *Intelligent Network*, [Q.1201]) geboren wurde, ist man einen anderen Weg gegangen, den Weg der stärkeren Zentralisierung von Dienstlogik und Dienstdaten.

Der Auslöser für die Entwicklung des IN-Konzeptes war der sogenannte *Freephone* Dienst, der in den USA als *800 service* und von der Deutschen Telekom als Service 130 bzw. ab 1998 als „freecall“ mit der Vorwahl 800 vermarktet wird. Die wesentliche Charakteristik dieses

20. Während im ISDN zwischen (Kommunikations-) Diensten im Sinne von Übermittlungsdiensten bzw. Telediensten einerseits und Zusatzdiensten bzw. zusätzlichen Dienstmerkmalen andererseits unterschieden wird, spricht man bei IN allgemein nur von Diensten. Darin sind dann sowohl die für die Teilnehmer sichtbaren Dienste (IN-Dienste), als auch „Dienste“ zum Verwalten und Erstellen dieser IN-Dienste enthalten. Die IN-Dienste sind dabei Mehrwertdienste, also Kommunikationsdienste, die um im IN realisierte Zusatzdienste erweitert werden. Erst in aktuelleren Dokumenten spricht die IN-Standardisierung von *telecommunication services*, welche bis dahin einfach nur *services* genannt wurden.

21. Der Begriff „Intelligentes Netz“ wird hier in Anführungszeichen gesetzt, da er mißverständlich ist. IN ist kein eigenständiges Netz. Es stellt eine Erweiterung bestehender Netze dar. Außerdem fügt es diesen keine „Intelligenz“ hinzu, sondern ermöglicht lediglich eine flexible Steuerung.

Dienstes ist es, daß der Dienstkunde unter einer einheitlichen und unter Umständen werbewirksamen²² Nummer für den Anrufer kostenlos erreichbar ist. Dabei kann das Rufziel ursprungs- und zeitabhängig vorgegeben werden. Dieser Dienst wird durch eine Rufnummerumsetzung realisiert. Erste Implementierungen führen die Rufe mit Hilfe der Vorwahl zu einer Sondervermittlungsstelle, in der die zugehörige Dienstlogik und die zugehörigen Dienstdaten verfügbar sind und die den Ruf dann zum eigentlichen Ziel weiterleitet. Dabei wird also eine zentrale Instanz in das Netz eingebracht, welche die Verwaltung von Dienstdaten sowie die Änderung und Erweiterung von Dienstlogik einfach hält. Diese Vorteile der Zentralisierung werden jedoch durch Mehraufwand bei der Nutzkanalführung und Grenzen bei der Skalierbarkeit erkauft. So muß jeder Ruf, der diesen Dienst beansprucht, über die Sondervermittlungsstelle geführt werden. Auf dem Hintergrund dieser Problematik wurde bei Bell Communications Research (Bellcore) 1984 das Konzept des „Intelligenten Netzes“ entwickelt.

Die wesentliche Idee des „Intelligenten Netzes“ ist die Trennung von Vermittlungsfunktionen, Dienstlogik und Dienstdaten sowie die Definition von Schnittstellen zwischen diesen Funktionalitäten. Die Kommunikation zwischen Instanzen an unterschiedlichen Stellen im Netz wird durch die nutzkanalunabhängige Signalisierung des Signalisiersystems Nummer 7 unterstützt. Die funktionale Trennung ermöglicht es beispielsweise, Dienstlogik und Dienstdaten zentral in sogenannten Dienststeuerknoten (*Service Control Points*, SCP) zu halten, während die Vermittlung der Rufe in geeigneten Vermittlungsstellen erfolgt, den sogenannten Dienstzugangsknoten (*Service Switching Points*, SSP). Diese Art der Zentralisierung ist typisch für derzeitige IN-Implementierungen, sie ist jedoch keine prinzipielle Eigenschaft von „Intelligenten Netzen“. Vielmehr bilden die Instanzen des IN eine dem Signalisier- und Nutzkanalnetz überlagerte Steuerungsstruktur, in der die physikalische Platzierung der Funktionalitäten relativ unabhängig von den darunterliegenden Netzstrukturen erfolgen kann. So bietet sich prinzipiell die Möglichkeit, die Dienststeuerung und die Dienstdatenhaltung auf verteilte Weise zu realisieren.

Die Motivation für die Entwicklung des Konzept des „Intelligenten Netzes“ ergibt sich aus dem Wunsch, eine Plattform für eine schnelle, kostengünstige und flexible Einführung und Verwaltung von neuen Diensten zu schaffen. Daraus ergeben sich für die Architektur des IN die Zielvorgaben der Dienst-, der Netz- und der Herstellerunabhängigkeit. Dienstunabhängigkeit²³ heißt, daß generische Schnittstellen und Funktionalitäten bereitgestellt werden, die es ermöglichen, innerhalb eines gewissen Rahmens eine Vielzahl von zuvor nicht näher spezifi-

22. So verwendete im Jahre 1996 die Deutsche Telekom AG die Nummer 0130 1996 für ein Informationssystem zu ihrem Börsengang. Sind Buchstaben des Alphabets Wählziffern zugeordnet, dann wird Werbung mit Nummern wie 1 800 PIZZA HUT oder 0130 FLEUROPE möglich. Solche Nummern werden im allgemeinen „vanity numbers“ genannt.

23. GSM wird gerne als das erste „intelligente“ Netz bezeichnet, da auch dort eine Trennung von Vermittlungsfunktionen in Mobilfunk-Vermittlungsstellen einerseits und Dienstlogik bzw. Dienstdaten (etwa zur Mobilitätsverwaltung) in den zugehörigen Steuerknoten (HLR/VLR) andererseits vorgenommen ist. Im Gegensatz zu IN sind die hier spezifizierten Schnittstellen und Funktionalitäten jedoch nicht dienstunabhängig, sondern, im Gegenteil, GSM-spezifisch.

zierten Diensten zu realisieren. Mit der Forderung nach Netzunabhängigkeit wird ein Anspruch auf Einsetzbarkeit für beliebige Kommunikationsnetze erhoben. Aktuelle Standards beschränken sich jedoch im wesentlichen auf PSTNs, ISDNs und PLMNs (*Public Land Mobile Networks*). Herstellerunabhängigkeit schließlich soll durch offene Standards erreicht werden und das Zusammenwirken von Anlagen verschiedener Hersteller ermöglichen. In den folgenden Abschnitten soll das IN-Konzept etwas detaillierter dargestellt werden und schließlich seine Anwendung zur Realisierung der Mobilitätsverwaltung diskutiert werden. Weiteres zur Einführung und Vertiefung von „Intelligenten Netzen“ findet sich in [GRKK93, DV92, KM94, Thö94, Baf95, MPZ96, FGKS97].

2.4.1 Das konzeptionelle Modell des „Intelligenten Netzes“

Das konzeptionelle Modell des „Intelligenten Netzes“ (*Intelligent Network Conceptual Model, INCM*) unterstützt die Beschreibung der IN-Architektur, indem es vier verschiedene Abstraktionsebenen definiert.²⁴ Auf jeder dieser Abstraktionsebenen wird die Architektur unter anderen Gesichtspunkten betrachtet. Es stellt damit einen Rahmen zur Beschreibung der Architektur und keinesfalls die Architektur selbst dar. Die vier Ebenen sind die Dienstebene, die globale funktionale Ebene, die verteilte funktionale Ebene und die physikalische Ebene (*Service Plane, Global Functional Plane, Distributed Functional Plane* und *Physical Plane*). In der ITU-T Standardisierung wird dieses Modell zur Gliederung verwendet. Die Mehrzahl der Einzelstandards beschränkt sich auf jeweils eine der Ebenen.²⁵ Abbildung 2-10 zeigt das konzeptionelle Modell des „Intelligenten Netzes“ in ähnlicher Weise wie es in [Q.1201] dargestellt wird.

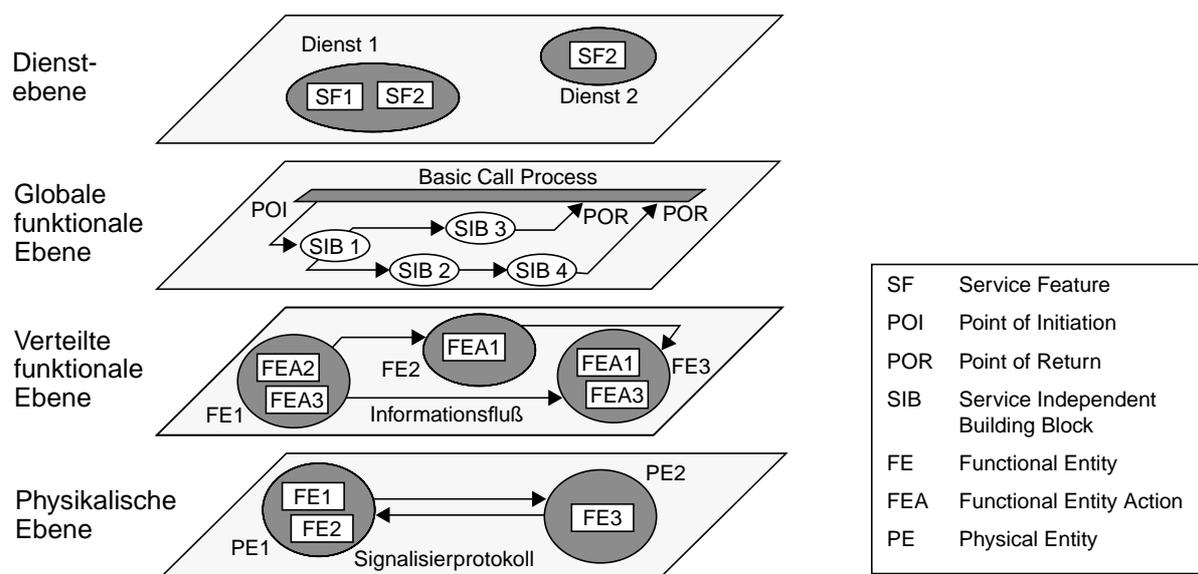


Abbildung 2-10: Das konzeptionelle Modell des „Intelligenten Netzes“.

24. Die Methodik ist eine Erweiterung des in der ITU-T Empfehlung I.130 (1988, *Method for the characterization of telecommunication services supported by an ISDN and network capabilities of an ISDN*) vorgestellten Dreistufenmodells zur Charakterisierung von Diensten.

25. Siehe Kapitel 2.4.2.

In der Dienstebene findet eine rein dienstorientierte Beschreibung statt, d.h. das Verhalten des Netzes dem Benutzer gegenüber wird betrachtet. Die eigentliche Implementierung ist irrelevant. Ein IN-Dienst wird durch eine Menge von Dienstmerkmalen (*Service Features*, SF) beschrieben. Ein Dienstmerkmal ist eine eigenständige Grundfähigkeit eines Dienstes, die nicht mehr sinnvoll zerlegbar ist, wie zum Beispiel Kurzwahl, Rufweiterleitung, Kostensplitting oder zeitabhängige Rufzielbestimmung. Dienste auf diese Weise zu beschreiben kann helfen, die Anforderungen an die darunterliegenden Ebenen zu ermitteln. Darüberhinaus kann in der Dienstebene das (erwünschte oder unerwünschte) Zusammenspiel mehrere Dienste oder Dienstmerkmale (*service interaction* [KK98]) untersucht und spezifiziert werden.²⁶

In der globalen funktionalen Ebene wird IN-spezifische Dienstlogik als Ganzes betrachtet. Die Dienstlogik wird mit Hilfe von dienstunabhängigen Bausteinen beschrieben, den sogenannten *Service Independent Building Blocks* (SIBs). SIBs repräsentieren (dienstunabhängige) Funktionalitäten, die im Netz als Ganzes vorhanden sind und durch deren Verkettung der Dienstablauf global beschrieben werden kann. Eine besondere Rolle spielt dabei der *Basic Call Process* (BCP)²⁷, der die Fähigkeit des Netzes, gewöhnliche Rufe zu behandeln und die Interaktion mit der IN-Dienstlogik bei deren Bearbeitung zu ermöglichen, repräsentiert. Die Beschreibung der Dienstlogik erfolgt durch die Spezifikation eines Auslösers im *Basic Call Process*, dem *Point of Initiation* (POI), und einer Verkettung von SIBs, die von einem POI ausgeht und sich in mehrere Teilzweige aufspalten kann, deren Enden über definierte Rückkehrpunkte (*Points of Return*, POR) zum BCP zurückgeführt werden. Nach solch einem POR erfolgt die gesamte Rufbehandlung wieder im BCP, während zwischen POI und POR ein Wechselwirken zwischen IN-Dienstlogik und gewöhnlicher Rufbehandlung stattfinden kann.

Neben der reinen Beschreibung von Dienstlogik sollen die SIBs auch dazu dienen, Dienstlogik zu erstellen. Dazu sollen sie elementare Funktionalitäten bereitstellen, für verschiedene Dienste wiederverwertbar sein, stabile Schnittstellen haben, flexibel verkettbar und parametrisierbar sein sowie dienst-, netz- und herstellerunabhängig sein. Die Dienstlogikmodellierung der ITU hat einen starken Einfluß auf Dienststellungsprogramme, die im allgemeinen SIBs als Konstruktionselemente verwenden. Die Standardisierung verwendet die SIBs jedoch derzeit im wesentlichen dazu, Anforderungen an Funktionalität in der verteilten funktionalen Ebene zu ermitteln, indem sie beispielhafte Dienstmerkmale mit Hilfe elementarer SIBs darzustellen versucht. In [RH95] wird in Bezug auf die Dienststellung festgestellt, daß das SIB-Konzept der ITU im Capability Set 1 von 1992 „noch nicht schlüssig, ausgereift und implementierbar ist“. Inzwischen wurde das Modell erweitert, so daß das Zusammenfassen mehrerer SIBs zu *high level SIBs* (HLSIB) und das parallele Ausführen von Dienstlogikprozessen dar-

26. Das Problem des Zusammenwirkens mehrerer Dienstmerkmale ist sehr komplex. Es kann nicht ausschließlich auf der Dienstebene betrachtet werden. Es müssen auch Aspekte der Implementierung der Dienste und der zugrundeliegenden Protokolle berücksichtigt werden (d.h. in den funktionalen und in der physikalischen Ebene).

27. Auch beim BCP handelt es sich um einen SIB.

stellbar wird. Insgesamt kann die Standardisierung hier jedoch nur einen Rahmen für Dienstleistungsprogramme bieten, in denen ausgereifere Konzepte der Softwaretechnik zur Anwendung kommen müssen.

Die Diskussion im vorigen Abschnitt illustriert, daß die internationale Standardisierung in den oberen beiden Ebenen des INCM Grenzen und Schwächen hat. In der verteilten funktionalen Ebene dagegen liegt der Schwerpunkt und die Stärke der IN-Standardisierung. In ihr wird der Tatsache Rechnung getragen, daß die Dienstleistung in verteilter Weise geschieht. Es werden sogenannte funktionale Instanzen (*Functional Entities*, FE) sowie Beziehungen zwischen diesen Instanzen festgelegt. Eine funktionale Instanz vereint und kapselt eine Gruppe von zur Dienstleistung notwendigen Funktionalitäten. Für die funktionalen Instanzen sind Aufgaben und Schnittstellen standardisiert. Zur Erbringung eines Dienstes werden „Aktionen“ in den funktionalen Instanzen durchgeführt (*Functional Entity Actions*, FEA) und Informationsflußkomponenten²⁸ über die Schnittstellen ausgetauscht (*information flow*). Es sind beispielsweise spezielle funktionale Instanzen zur Rufsteuerung, zur Dienststeuerung, zur Dienstdatenhaltung und zum Erbringen besonderer Dienstmerkmale definiert.

Während in der verteilten funktionalen Ebene die prinzipielle Möglichkeit der Verteilung der Funktionalitäten betrachtet wird, werden in der physikalischen Ebene die konkreten Möglichkeiten der Verteilung spezifiziert. Dazu werden verschiedene Typen von physikalischen Instanzen²⁹ (*Physical Entities*) mit Hilfe der in ihnen lokalisierten funktionalen Instanzen definiert und die entsprechenden Protokolle zur Realisierung des Informationsflusses zwischen funktionalen Instanzen in verschiedenen physikalischen Instanzen benannt. Eine einzelne funktionale Instanz kann nicht auf mehrere physikalische Instanzen verteilt werden. Eine Informationsflußkomponente in der verteilten funktionalen Ebene wird in der Regel auf eine Operation des *Intelligent Network Application Part* (INAP) in der physikalischen Ebene abgebildet. Neben den schon eingeführten Dienststeuerknoten (SCP) und Dienstzugriffsknoten (SSP) sind zur Zeit vor allem die Dienstdatenknoten (*Service Data Points*, SDP) und die intel-

28. In den ITU-T Empfehlungen wird das Anstoßen einer FEA in einer funktionalen Instanz durch eine andere funktionale Instanz „*information flow*“ genannt. Eine Auflistung aller möglichen Optionen, FEAs in anderen funktionalen Instanzen anzustoßen, wird entsprechend als Auflistung von „*information flows*“ bezeichnet. Diese etwas unschöne Verwendung von Singular und Plural des Wortes *flow* wird hier nicht ins Deutsche übernommen. Stattdessen wird die Gesamtheit des Informationsaustausches als Informationsfluß bezeichnet. Ein einzelner sogenannter „*information flow*“ wird als Informationsflußkomponente bezeichnet. Demgemäß kann man dann von einer Auflistung von Informationsflußkomponenten („*information flows*“) sprechen.

29. Zur Übersetzung: In Entwürfen zur ITG-Empfehlung „Systeme der Vermittlungstechnik“ (Teil des Begriffswerks) wurde einige Zeit der Begriff *physikalische Elemente des Intelligenten Netzes* verwendet, der aber in neueren Empfehlungen nicht mehr vorkommt. Bei den bisher verabschiedeten Normen zum „Intelligenten Netz“ befinden sich alle physikalischen Instanzen im Einflußbereich von Netzbetreibern. Man könnte sie deshalb auch als *Netzknoten* bezeichnen. Wird jedoch beispielsweise Dienstlogik in Endgeräte ausgelagert, so ist die Verwendung des Begriffs *Netzknoten* für Endgeräte etwas unglücklich, während der Begriff *physikalische Instanz* auch diese umfaßt.

l intelligenten Peripherieeinrichtungen (*Intelligent Peripherals*, IP), die zur Erbringung besonderer Dienstmerkmale – wie etwa der Spracherkennung oder dem Abspielen von Ansagen – dienen, von Bedeutung.

Die Modellkomponenten der verschiedenen Ebenen des INCM stehen in einer gewissen Beziehung zueinander. Idealerweise lassen sich Abbildungen zwischen den Komponenten der verschiedenen Ebenen finden. So kann ein Dienst oder ein Dienstmerkmal in der Dienstebene durch die Beschreibung der Dienstlogik, das heißt durch eine Verkettung von SIBs, in der globalen funktionalen Ebene dargestellt werden. Die Ausführung eines SIBs wiederum läßt sich in der verteilten funktionalen Ebene durch „Aktionen“ in den funktionalen Instanzen und durch Informationsfluß repräsentieren, was wiederum auf Bearbeitungsprozesse und den Austausch von Signalisierungsmeldungen in der physikalischen Ebene abgebildet werden kann. In [Baf95, BS95, BS96] haben wir gezeigt, wie sich mit Hilfe solcher Abbildungen wichtige Daten zur Kapazitätsplanung und Leistungsuntersuchung für Netze mit IN-basierter Dienstleistung gewinnen lassen.

2.4.2 Internationale Standardisierung des „Intelligenten Netzes“

Die ersten proprietären IN-Konzepte wurden in den achtziger Jahren von Bellcore vorgelegt. IN/1 diente zur Rufnummerumsetzung, zum Beispiel im Rahmen des *Freephone* Dienstes, und fand in den USA große Verbreitung. IN/2 sollte dem nachfolgen, wurde jedoch nach der Definition der Zwischenstufe IN/1+ zugunsten eines neuen Konzepts, dem 1989 vorgestellten *Advanced Intelligent Network* (AIN), aufgegeben. Basierend auf denselben Konzepten wie AIN begann 1989 auch die internationale und europäische Standardisierung bei der ITU-T beziehungsweise bei ETSI, hier nun wieder unter Verwendung des Begriffs *Intelligent Network*. Trotz mancher Unterschiede im Detail sind sich AIN und ITU-T IN konzeptionell sehr ähnlich. Die Standardisierung bei ETSI erfolgt in Übereinstimmung mit derjenigen bei der ITU-T. Somit ist es ausreichend, im folgenden nur noch die Standardisierung der ITU-T vorzustellen.

Die Standardisierung des IN bei der ITU-T erfolgt in Phasen. Das Ziel ist eine Architektur in der beliebige IN-Dienste effizient erbracht sowie schnell und einfach entwickelt, eingeführt und verwaltet werden können. Auf dem Weg zu solch einer Zielarchitektur werden als Zwischenschritte sogenannte *Capability Sets* verabschiedet, die einen gewissen Funktionsumfang im IN spezifizieren. Je nach *Capability Set* sind dann mehr oder weniger mächtige Protokolle und Schnittstellen zur Dienstleistung, Dienstverwaltung und Dienstleistung für unterschiedliche Netztypen definiert. Das 1992 verabschiedete *Capability Set 1* (CS-1) konzentriert sich auf die Dienstleistung von relativ einfachen, halbrufbezogenen³⁰ Diensten in klassischen Telekommunikationsnetzen (PSTN, ISDN, PLMN). Die 1997 verabschiedeten Empfehlungen des *Capability Set 2* (CS-2) tragen im Vergleich dazu der Dienstleistung, dem Dienstmanagement und neuen Anforderungen bezüglich Teilnehmermobilität in stärkerem

30. Übersetzung von *single-ended*. Siehe dazu S. 34.

Maße Rechnung. Hierzu gehören zum Beispiel rufunabhängige Dienstausröser, die für die Registrierung von Teilnehmern an Endgeräten verwendet werden können. Die Arbeit am *Capability Set 3* (CS-3) hat im Jahre 1995 begonnen, mit dem Ziel Endgerätemobilität, B-ISDN und Multimedia zu unterstützen sowie Diensterstellung und Dienstmanagement zu standardisieren. Vor dem Jahr 2000 ist mit dem Abschluß der Arbeiten am CS-3 kaum zu rechnen.

Abbildung 2-11 gibt eine Übersicht über die Struktur der Empfehlungen der ITU-T zum „Intelligenten Netz“. Die prinzipielle Architektur des „Intelligenten Netzes“ ist in den Empfehlungen Q.120x festgelegt. Die Empfehlungen Q.121x beschreiben das CS-1. Entsprechend beschreiben die Empfehlungen Q.122x das CS-2 und so fort. Innerhalb einer solchen Serie steht die letzte Ziffer der Empfehlung für einen bestimmten Teilaspekt. Die Empfehlungen mit den Ziffern 2, 3, 4 und 5 am Ende beispielsweise befassen sich gemäß dem INCM mit der Dienstebene, der globalen funktionalen Ebene, der verteilten funktionalen Ebene beziehungsweise der physikalischen Ebene. Die Empfehlung Q.1290 gibt einen Überblick über die verwendete Terminologie. In den folgenden beiden Unterkapiteln wird zuerst das *Capability Set 1* detaillierter vorgestellt und danach auf die *Capability Sets 2* und *3* eingegangen.

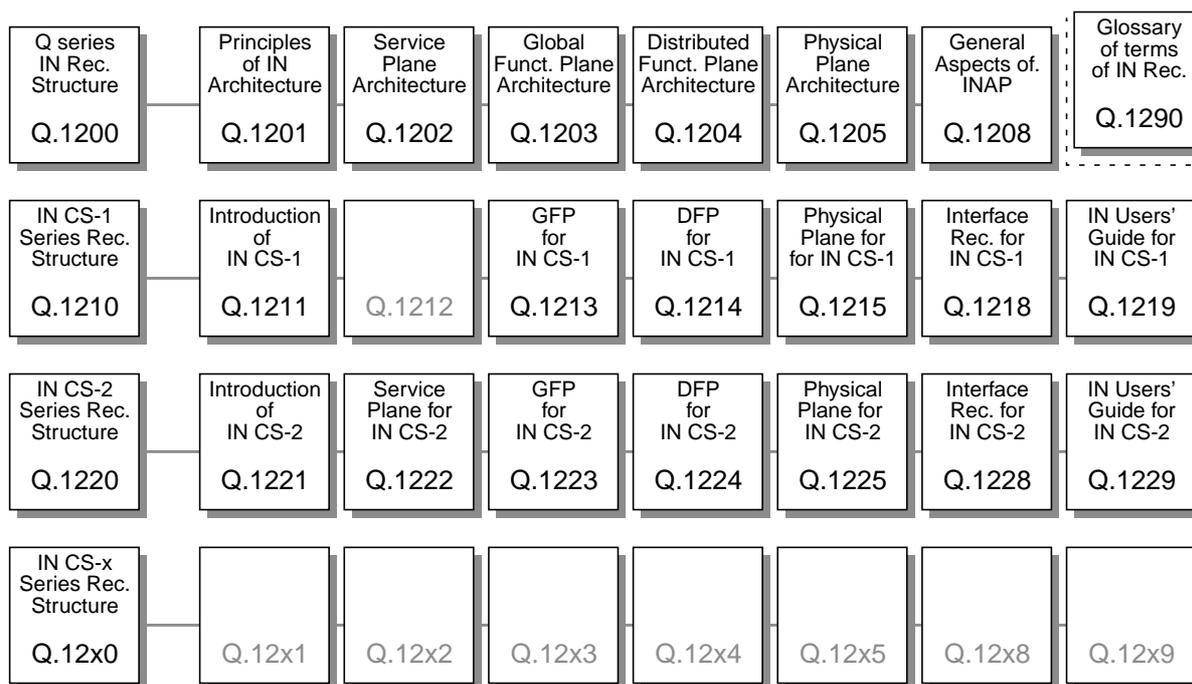


Abbildung 2-11: Übersicht über die Struktur der ITU-T IN-Empfehlungen
(nach [MPZ96, Q.1200]).

2.4.2.1 Das ITU-T IN Capability Set 1

Das *Capability Set 1* (CS-1) stellt den ersten Schritt der Standardisierung auf dem Weg zur IN-Zielarchitektur dar. Man beschränkt sich hier auf die Dienstleistung in klassischen Telekommunikationsnetzen und auf eine relativ einfache Funktionalität. Dienste, die innerhalb des CS-1 realisiert werden können,³¹ lassen sich gemäß [MPZ96] im wesentlichen in drei³² Kategorien einteilen: Dienste zur Rufnummernumsetzung (z.B. Kurzwahl oder bundeseinheitliche

Rufnummer), Dienste mit besonderer Entgelterfassung (z.B. Kreditgespräch oder Kostensplitting) und Dienste zum Ausfiltern von Rufen (z.B. Einschränkung von abgehenden Rufen oder Einschränkung von kommenden Rufen). Entsprechend werden im CS-1 dreizehn relativ einfache Dienstbausteine (SIBs) vorgeschlagen. Davon seien genannt: *Charge* – zur Beeinflussung der Entgelterfassung, *Log Call Information* – zum Aufzeichnen von Rufdaten, *Screen* – zum Ausfiltern von Rufen, *Translate* – zur Nummernübersetzung sowie *User Interaction* – zur Interaktion mit Teilnehmern.

Die Beschränkungen bezüglich des Funktionsumfangs auf den oberen beiden Ebenen des INCM finden sich auch in den unteren beiden Ebenen wieder. Zur Diensterbringung sind vier IN-spezifische funktionale Instanzen definiert zwischen denen drei Schnittstellen festgelegt sind. Die IN-spezifischen funktionalen Instanzen sind die Dienstzugangssteuerung (*Service Switching Function, SSF*), die Dienststeuerung (*Service Control Function, SCF*), die funktionale Instanz zur Verwaltung von Dienstdaten (*Service Data Function, SDF*) und die funktionale Instanz zum Erbringen besonderer Dienstmerkmale (*Specialized Resource Function, SRF*). Zwischen SCF einerseits und SSF, SRF sowie SDF andererseits sind die Schnittstellen D, E beziehungsweise F definiert. Zwischen SSF, SRF und SDF gibt es keine IN-spezifischen Schnittstellen.

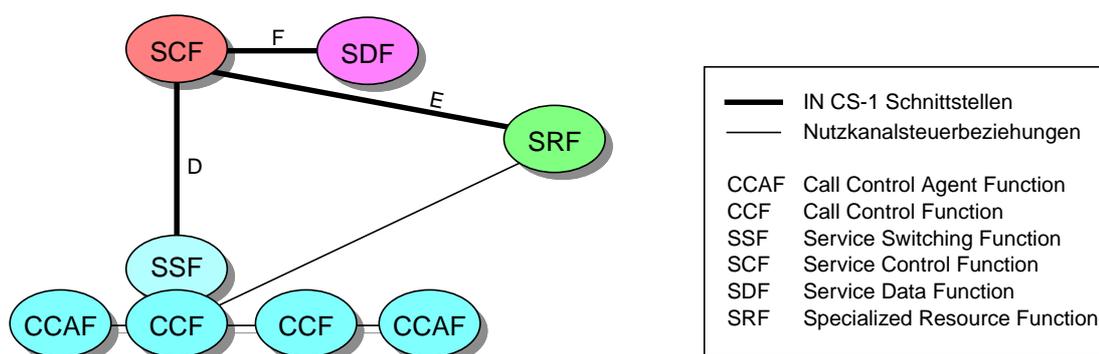


Abbildung 2-12: Funktionale Instanzen und zugehörige Schnittstellen im IN CS-1.

Abbildung 2-12 zeigt das Modell des IN CS-1 in der verteilten funktionalen Ebene.³³ Die dort außerdem noch dargestellten funktionalen Instanzen Rufsteuerungsagent (*Call Control Agent Function, CCAF*) und Rufsteuerung (*Call Control Function, CCF*) modellieren Funktionalität

31. Da das IN eine Plattform für neue Dienste ist, lassen sich nicht alle möglichen Dienste von vornherein klassifizieren. In den CS-1 Standards sind jedoch einige Referenzdienste angegeben, die mit der CS-1 Architektur realisierbar sein sollen. Auf diese Referenzdienste bezieht sich die hier angegebene Kategorisierung. Der Dienst *Einspielen von Informationen* (z.B. Begrüßung, Werbung, Gesprächstarif) vor dem Verbindungsaufbau ließe sich beispielsweise im CS-1 realisieren, fällt jedoch nicht unbedingt in eine der angegebenen Kategorien.
32. [MPZ96] schlägt außerdem noch die Kategorie „other services“ vor, in die z.B. *Televoting*, *Mass Calling* und *Malicious Call Identification* eingeordnet werden.
33. In Q.1211 [Q.1211] werden außerdem noch funktionale Elemente zum Dienstmanagement und zur Diensterstellung genannt, die jedoch in Q.1214 [Q.1214] nicht näher spezifiziert sind. Es handelt sich dabei um die *Service Management Function* (SMF), die *Service Management Access Function* (SMAF) und die *Service Creation Environment Function* (SCEF).

ten in Vermittlungsstellen, die weitgehend IN-unabhängig sind. Die CCAF repräsentiert dem Teilnehmer gegenüber die Rufsteuerung (CCF) und leitet dessen Anforderungen an die CCF weiter. Die Rufsteuerung (CCF) entspricht der klassischen Verbindungssteuerungsfunktion in einer Vermittlungsstelle. Für den Einsatz im IN muß sie so erweitert werden, daß sie sich im Zusammenspiel mit der Dienstzugangssteuerung (SSF) von der SCF steuern läßt. Dazu muß die Rufbehandlung angehalten, ein Dienst in der SCF ausgelöst, der SCF eine Sicht auf den Zustand des Rufes gegeben und auf Steuerkommandos von der SCF reagiert werden können. Die eigentliche Dienststeuerung findet dabei in der SCF statt, die bei Bedarf Dienstdaten aus der SDF bezieht oder diese dort verändert. Darüberhinaus kann ein Teilnehmer über einen Nutzkanal mit der SRF verbunden werden, die auf Anweisung der SCF besondere Dienstmerkmale erbringt und das Ergebnis ihrer Ausführung an die SCF zurückmeldet. Als Beispiel sei das Spielen einer Ansage und das Entgegennehmen von Steuerkommandos via DTMF oder Spracherkennung genannt.

Die IN CS-1 Dienste werden als *single-ended and single-point-of-control services* bezeichnet. Bei der Modellierung eines Rufes zwischen zwei Teilnehmern verwendet man zwei unabhängige Teilmodelle für den kommenden bzw. den gehenden Teil eines Rufes („Halbrufe“). Wirkt Dienstlogik direkt nur auf eines der beiden Teilmodelle ein, so nennt die ITU-T Standardisierung dies *single-ended* (halbrufbezogen). Liegt die Dienststeuerung bezüglich eines solchen Halbrufes zu jedem Zeitpunkt bei maximal einer SCF, so spricht man von einem *single point of control*. Das bedeutet, daß während der Phase des Rufes, in der die Dienstlogik aktiv ist, genau eine SCF einem Halbruf der SSF/CCF³⁴ zugeordnet sein kann. Die vorgeschlagene interne Struktur von SSF/CCF und die dazugehörigen Rufmodelle (*Basic Call State Models*, BCSMs) sind in Anhang B dargestellt. Sie geben einen Eindruck davon, auf welche Weise die Dienstlogik an Auslösern (*Detection Points*), die sich an Übergängen zwischen verschiedenen Rufphasen (*Points in Call*) befinden, Einfluß auf die Rufbearbeitung nehmen kann.

Abbildung 2-13 zeigt die wichtigsten physikalischen Instanzen im CS-1: SSP, SCP, SDP und IP.³⁵ Für die Signalisierung ist der *Intelligent Network Application Part* (INAP) definiert, der sich auf die *Transaction Capabilities* des Signalisiersystems Nummer 7 abstützt. Zwischen SSP und IP kann auch Signalisierung über den ISDN D-Kanal (Q.931) unter Verwendung von ROSE (X.219) erfolgen. Ist die intelligente Peripherieeinrichtung (IP) in diesem Fall nicht direkt über das Signalisiernetz erreichbar, so erfolgt der APDU-Austausch über den SSP. Prinzipiell kann die SCCP-gestützte IN-Signalisierung zwischen physikalischen Instanzen in verschiedenen Netzen stattfinden, so daß beispielsweise die Dienststeuerung durch den SCP eines fremden Netzbetreibers erfolgen kann. Bisherige Realisierungen und die Abschottung der netzbetreibereigenen Signalisiernetze unterstützten dies jedoch noch nicht.

34. Die Schnittstelle zwischen CCF und SSF ist nicht standardisiert. Sie wird als herstellerspezifisch betrachtet. Aus diesem Grund werden CCF und SSF oft als Einheit betrachtet (SSF/CCF).

35. Daneben sind noch definiert: *Adjunct*, *Service Node* und *Network Access Point*.

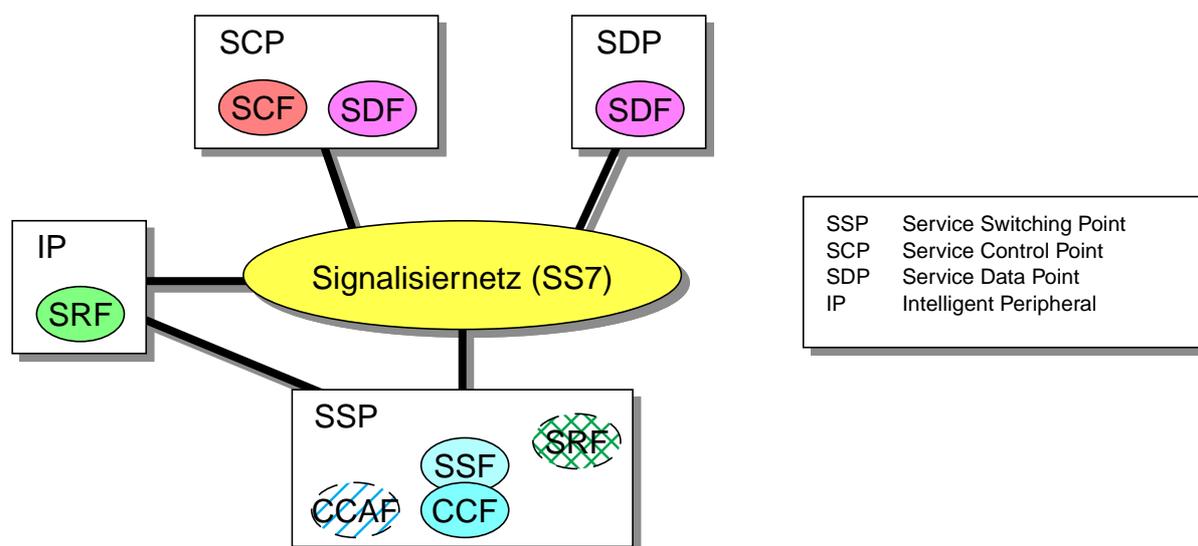


Abbildung 2-13: Die wichtigsten physikalischen Instanzen im IN CS-1 und die ihnen (z. T. optional) zugeordneten funktionalen Instanzen.

2.4.2.2 Weitere Capability Sets

Die zukünftigen Standards zum IN Capability Set 2 (CS-2) [Q.1220] wurden 1997 zur Verabschiedung vorgelegt. Wie schon im CS-1 werden ausschließlich *single-ended and single-point-of-control services* unterstützt. Diese Dienste werden im CS-2 erstmals näher bezeichnet. Sie werden in die Klasse der Kommunikationsdienste (*telecommunication services*) eingeordnet und damit nicht mehr lediglich *services* genannt. Außerdem werden zwei neue Klassen von „Diensten“ eingeführt: Dienstverwaltungsdienste (*service management services*) und Dienstleistungsdienste (*service creation services*). Diese sollen im weiteren nicht näher betrachtet werden, da zum einen die Standardisierung noch nicht weit genug fortgeschritten ist und zum anderen die eigentlichen Kommunikationsdienste für die vorliegende Arbeit von deutlich größerer Bedeutung sind.

In Q.1221 [Q.1221] werden einige Referenzdienste genannt, die sich im wesentlichen in drei Gruppen einteilen lassen:

- ❑ Typische CS-1 Dienste, die netzübergreifend erbracht werden. Beispiele sind: *Internetwork*³⁶ *Freephone*, *Internetwork Premium Rate*, *Internetwork Televoting* und *Global Virtual Network Service*³⁷.
- ❑ Anschlußbezogene Zusatzdienste.³⁸ Beispiele sind: Anklopfen (*Call Waiting*), Anrufweilerschaltung (*Call Transfer*) und Makeln (*Call Hold*).

36. Außerhalb der IN CS-2 Dokumente wird statt *Internetwork* üblicherweise der Begriff *International* verwendet. So nennt sich *Internetwork Freephone* beispielweise *International Freephone Service* (ITU-T Rec. E.152).

37. Es handelt sich hierbei um ein internationales VPN (Virtual Private Network).

38. Diese Dienste sind bereits in vielen modernen Ortsvermittlungstellen realisiert und werden als T-Net Komfortleistungsmerkmale (Deutsche Telekom), ISDN-Leistungsmerkmale oder als CLASS (*Custom Local Area Signaling Services*) Angebote (USA) vermarktet.

- UPT – Universelle Persönliche Telekommunikation. Ein Dienst, der personenbezogene Mobilität unterstützt. Wichtige Dienstmerkmale sind: Authentifizierung von rufendem und gerufenem Teilnehmer sowie Registrierung an beliebigen Endgeräten, so daß dort Rufe abgesetzt und entgegengenommen werden können.

Außerdem sind noch einige andere Dienste genannt, so zum Beispiel ein Dienst zum Versenden von Nachrichten (*Message Store and Forward*). Darüberhinaus soll Endgerätemobilität und multimediale Breitbandkommunikation unterstützt werden. Dies zielt auf B-ISDN und die dritte Generation von Mobilkommunikationsnetzen (UMTS und IMT-2000), die Teil eines zukünftigen B-ISDN sein sollen. Da jedoch die vorläufigen IN CS-2 Standards auf den darunterliegenden Ebenen des INCM im Kern noch nicht die zugehörigen Anforderungen berücksichtigen, werden die entsprechenden Dienste wohl erst in zukünftigen *Capability Sets* oder in der Erweiterung des IN CS-2 realisiert werden können.

Dienstmerkmale, die nicht unmittelbar mit einem Ruf zusammenhängen, spielen eine wichtige Rolle bei der Unterstützung von personenbezogener Mobilität. Dies spiegelt sich sowohl auf der globalen als auch auf der verteilten funktionalen Ebene des IN CS-2 wider, wo erstmals Instanzen zur rufunabhängigen Dienstleistung spezifiziert sind. In der globalen funktionalen Ebene wird neben dem *Basic Call Process* (BCP) nun auch ein *Basic Call-Unrelated Process* (BCUP) definiert. Überdies wird die funktionale Ebene um einige SIBs erweitert (u.a. *Authentication*). Neu sind die Elemente *SIB operation* und *High Level SIB* (HLSIB), die eine Verfeinerung beziehungsweise Zusammenfassung von SIBs ermöglichen. Schließlich sind parallel verlaufende Dienstlogikprozesse vorgesehen.

Auf der verteilten funktionalen Ebene [Q.1224] gibt es drei neue funktionale Instanzen, wovon zwei zur rufunabhängigen Dienstleistung beitragen. In Analogie zum Rufsteuerungsagenten (CCAF) wird der Dienstlogikagent (*Service Control User Agent*, SCUAF) definiert, der einem Teilnehmer den rufunabhängigen Zugang zu Dienstlogikinstanzen ermöglicht. Analog zur (rufbezogenen) Dienstleistungssteuerung (SSF) wird die rufunabhängige Dienstleistungssteuerung (*Call-Unrelated Service Function*, CUSF) eingeführt, welche die Beziehungen zwischen SCAF und SCF verwaltet. Sie ist beispielsweise für das Auslösen von SCF-Dienstlogik zuständig und kann in ihrem Verhalten von der SCF beeinflußt werden. Neu ist außerdem noch die IN-Zugangsfunktion (*Intelligent Access Funktion*, IAF), die sich in Knoten ohne IN-Funktionalität (einfachen Datenbanken, Nebenstellenanlagen oder Endgeräten) befinden kann und eine Schnittstelle zur SCF ermöglichen soll.

Abbildung 2-14 zeigt die funktionalen Instanzen der verteilten funktionalen Ebene des IN CS-2 sowie die zugehörigen Schnittstellen zwischen diesen.³⁹ Im Vergleich zum IN CS-1 hat sich die Zahl der IN-spezifischen Schnittstellen von drei auf acht erhöht. Dazu gehören die Schnittstellen zwischen den neuen funktionalen Instanzen CUSF sowie IAF einerseits und SCF andererseits. Neu sind auch Schnittstellen zwischen SCFs (SCF-SCF) bzw. zwischen

39. Wie schon bei der Darstellung zum IN CS-1 sind SMF, SMAF und SCEF sowie die zugehörigen Schnittstellen nicht dargestellt.

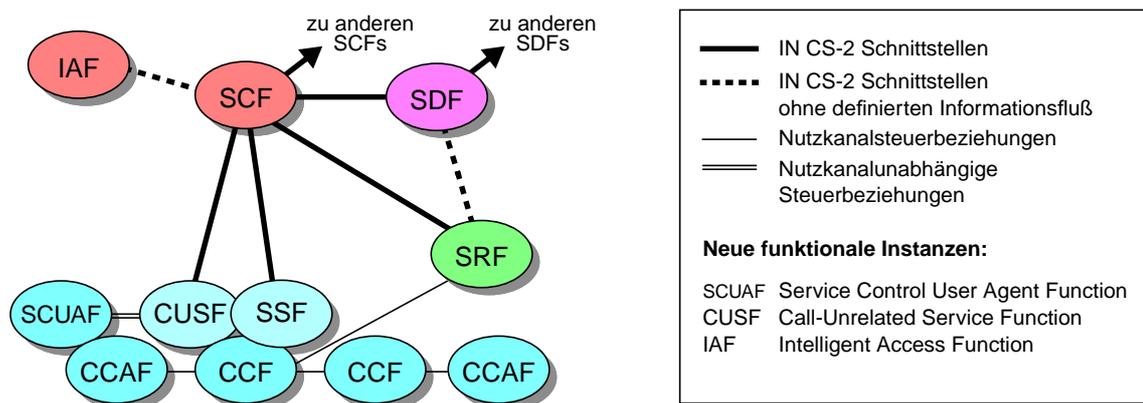


Abbildung 2-14: Funktionale Instanzen* und zugehörige Schnittstellen im IN CS-2.

* Nicht aufgeführte Akronyme sind in Abbildung 2-12 erläutert.

SDFs (SDF-SDF) sowie zwischen SRF und SDF (SRF-SDF). Wobei jedoch für SCF-IAF und SRF-SDF noch kein genauer Informationsfluß definiert ist. Bei der SCF-IAF-Beziehung spielt die IAF voraussichtlich die Rolle einer *unterstützenden SCF*, wie es in einer SCF-SCF-Beziehung vorgesehen ist. Schnittstellen zwischen SCUAF und CCAF sowie zwischen CUSF und SSF/CCF sind (wie schon bei SSF/CCF) nicht näher spezifiziert. Die Schnittstelle zwischen SCUAF und CUSF ist keine IN-spezifische Schnittstelle (analog zu CCAF-CCF).

Die wichtigen Neuerungen auf der verteilten Funktionalen Ebene sind also die rufunabhängige Dienstbringung und der mögliche Informationsfluß zwischen SDFs bzw. SCFs. Außerdem wurden den Rufmodellen (BCSMs) weitere *Points in Call* und weitere *Detection Points* hinzugefügt.

Auf der physikalischen Ebene (Abbildung 2-15) wird der SSP um die CUSF erweitert. Auch SCPs bzw. SDPs kommunizieren untereinander über das Signalisiersystem Nummer 7. Neu – und erst spät in die Standardisierung aufgenommen – ist ein rufunabhängiger Dienstzugangsknoten (*Call-Unrelated Service Point, CUSP*), der im wesentlichen die CUSF enthält. Der Dienstlogikagent (SCUAF), der dem Teilnehmer den Zugang zur CUSF und damit zur rufunabhängigen Dienstbringung ermöglicht, soll sich in Geräten beim Teilnehmer – in sogenanntem *Customer Premises Equipment (CPE)* – befinden. Damit wird erstmals eine IN-spezifische funktionale Instanz beim Teilnehmer plazierte. Hier deutet sich an, wie bei immer „intelligenten“ Endgeräten die Dienstbringung nicht mehr nur ausschließlich im Netz stattfinden muß.

Anhang A der Empfehlung Q.1224, der Mobilitätsaspekte der verteilten funktionalen Ebene behandelt, gibt eine Vorstellung davon, wohin sich die IN-Standardisierung mit dem *Capability Set 3 (CS-3)* entwickeln könnte. Dort wird gezeigt, wie das Modell der funktionalen Ebene um weitere funktionale Instanzen und Schnittstellen zur Verwaltung von Funkverbindungen zu mobilen Endgeräten erweitert werden kann. Abbildung 2-16 zeigt das zugehörige funktionale Modell. Besonders zu nennen sind die *Call-Unrelated Radio Access Control Function (CURACF)* und die *Call-Related Radio Access Control Function (CRACF)*. Neu ist außerdem die *Radio Control Function (RCF)*.

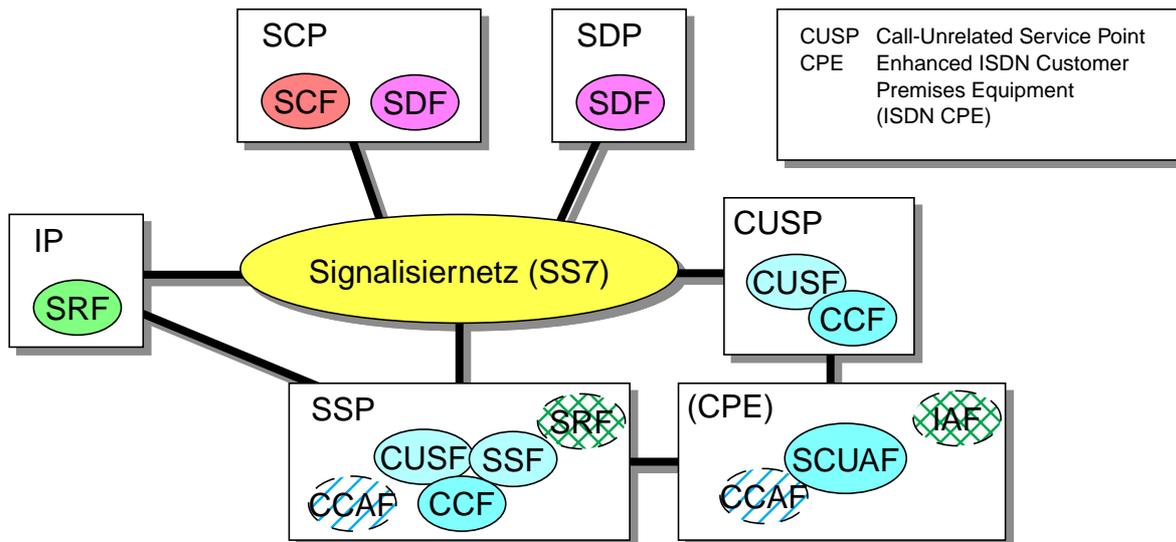


Abbildung 2-15: Die wichtigsten physikalischen Instanzen im IN CS-2 und die ihnen (z. T. optional) zugeordneten funktionalen Instanzen.

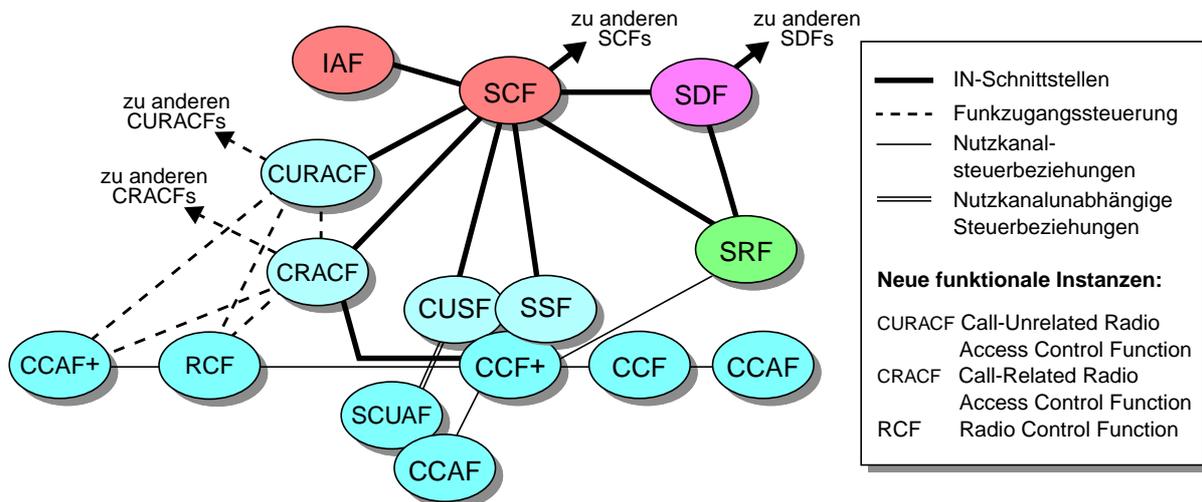


Abbildung 2-16: Vorschlag zur Unterstützung von Endgerätemobilität in der funktionalen verteilten Ebene (gemäß Anhang A von Q.1224).

Bei der Standardisierung (ETSI) und in der Forschung (RACE, ACTS) zu UMTS, der dritten Generation von Mobilkommunikationssystemen in Europa, wird davon ausgegangen, daß wichtige Steuerfunktionen im Rahmen des „Intelligenten Netzes“ realisiert werden [Buit94, CH94, FBWC97]. Dazu gehörte bisher auch die Mobilitätsverwaltung. In [ETSI96] wird das IN/UMTS „generische funktionale Modell“ von ETSI vorgestellt (s. Abbildung 2-17), das ein IN-Modell in der verteilten funktionalen Ebene ist. Obwohl es inzwischen Bestrebungen gibt, die Mobilitätsverwaltung auf Basis des MAP von GSM zu entwickeln [Dav97], so zeigt das Modell dennoch sehr gut, wohin sich die IN-Standardisierung entwickeln könnte. Drei Besonderheiten sind zu nennen. Wie auch beim CS-2 ist rufunabhängige Dienstleistung möglich, hier via die *Service Control Agent Function* (SCAF). Neben den Instanzen zur Rufsteuerung sind auch Instanzen zur Verbindungssteuerung vorgesehen, so zum Beispiel die *Bearer Control Function* (BCF) und die *Radio Bearer Control Function* (RBCF). Darüberhinaus sind erstmals

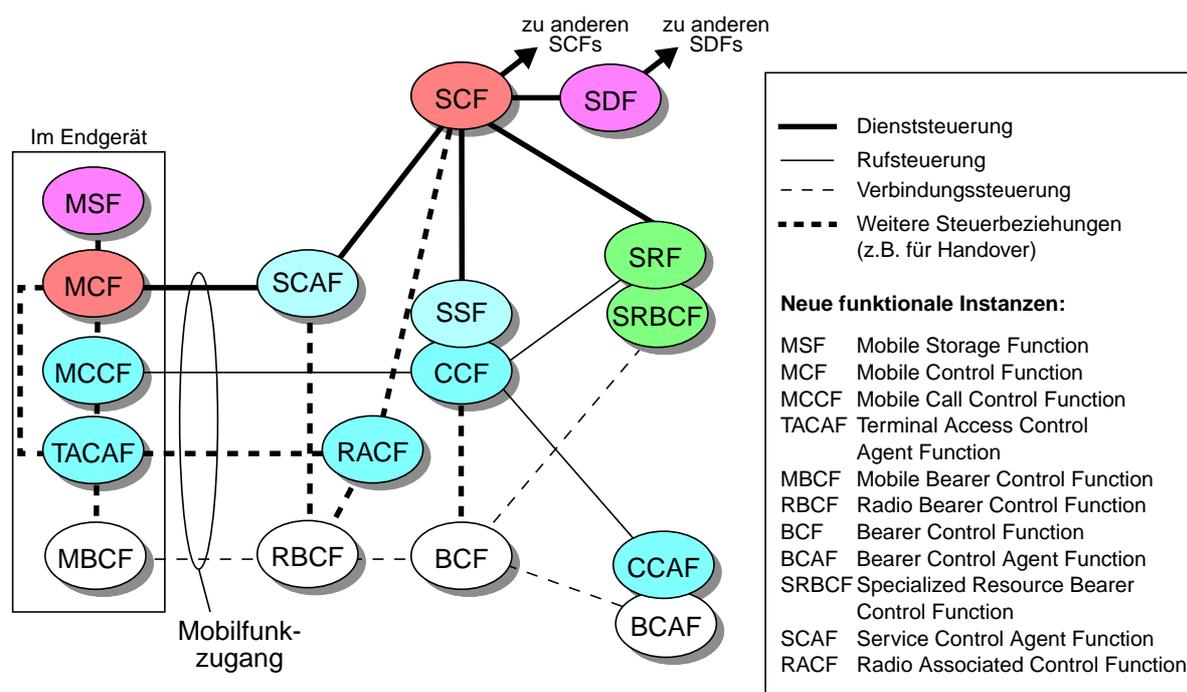


Abbildung 2-17: Generisches funktionales IN-Modell für UMTS (nach [ETSI96]).

konkrete funktionale Instanzen im Endgerät definiert: Die *Mobile Storage Function* (MSF), eine Art SDF für mobilitätsbezogene Daten im Endgerät, die *Mobile Control Function* (MCF), die – als Pendant zur SCF – Dienstlogik zur Mobilitätsverwaltung im Endgerät beinhaltet, sowie die MCCF, die TACAF und die MCBF zur Unterstützung von Ruf-, Handover- und Verbindungssteuerung.

Abschließend kann man also sagen, daß zukünftige IN-Standards verstärkt rufunabhängige Dienstleistung, Mobilität – sowohl von Personen als auch von Endgeräten – und, durch die Trennung von Ruf- und Verbindungssteuerung, Multimedia-Dienste und Dienste mit mehreren Beteiligten unterstützen werden.

2.4.3 TINA

Das Konzept des „Intelligenten Netzes“ stammt aus den frühen achtziger Jahren. Neuerungen in der Softwaretechnik wie Standards für offene, verteilte Anwendungen im Rahmen des *Open Distributed Processing*⁴⁰ (ODP) eröffnen neue Möglichkeiten für die langfristige Entwicklung des IN. Bei Bellcore wurde 1990 begonnen, ein neues Architekturkonzept für die Erbringung von Telekommunikationsdiensten zu erarbeiten, die *Information Networking Architecture* (INA). 1992 wurde ein internationales Konsortium gegründet, das, beeinflusst von INA, eine gemeinsame, offene Software-Architektur für Multimedia- und Informationsdienste definiert

40. ITU-T Rec. X.900 Serie bzw. ISO/IEC 10746 Serie. Bisher erschienen: X.901 [X.901], X.902 [X.902] und X.903 [X.903] bzw. ISO/IEC 10746-1, ISO/IEC 10746-2 und ISO/IEC 10746-3 sowie ISO/IEC DIS 10746-4.

[DH97, MM96]. Das Architekturkonzept wird als *Telecommunications Information Networking Architecture* (TINA), das Konsortium ist als *TINA-Consortium* (TINA-C, [DNI95]) bezeichnet.⁴¹

Die weitere Entwicklung des „Intelligenten Netzes“ wird sicherlich durch TINA beeinflusst werden [DK95]. Inwieweit eine Ablösung von IN durch TINA beziehungsweise eine Integration von IN und TINA erfolgen wird, ist schwer vorhersagbar, da die beiden Konzepte trotz mancher Gemeinsamkeiten auch große Unterschiede aufweisen. Die Motivation und die Zielvorgaben für TINA entsprechen im wesentlichen denen von IN.⁴² Während IN aber darauf abzielt, bestehende Telekommunikationsnetze um neue Dienste bzw. Dienstmerkmale zu erweitern, wird bei TINA eine völlig neue Architektur definiert, die das gesamte Spektrum von Signalisierung über Diensterbringung bis zu Netzmanagement abdeckt und dabei Funktionalität in bestehenden Telekommunikationsnetzen entweder nicht berücksichtigt oder einer unabhängig definierten, abstrakten Sicht unterordnet. Dieser Ansatz hat es ermöglicht, TINA unabhängig von Ballast aus mehr als einhundert Jahren Telefonie – praktisch „auf der grünen Wiese“ – zu definieren. Daraus resultiert jedoch andererseits das Problem, daß eine Evolution bzw. eine Migration von bestehender Telekommunikationstechnik zu TINA nur teilweise und nur langsam vollzogen werden kann [LMMS97]. Heutige Numerierung und Adressierung, Endgeräte, Teilnehmerschnittstellen, Signalisierungsprotokolle, Vermittlungsstellen, Rufsteuerungsprogramme, Netzmanagementsysteme, Dienststeuerknoten, Mobilitätsverwaltungsprotokolle und Entgelterfassungssysteme sind im Grunde inkompatibel zu TINA. Erst eine Revolution der Netztechnik könnte TINA eine Chance, von Anfang an voll zur Entfaltung zu kommen, bieten.

Nichtsdestoweniger gibt TINA wichtige Impulse in Bezug auf das Zusammenwachsen von Informationstechnik (IT) und Telekommunikationstechnik. Im Gegensatz zum heutigen IN, wo einfache Client-Server-Beziehungen vorherrschen, stützt sich TINA auf moderne, objektorientierte Konzepte für verteilte Systeme ab. Das *TINA Distributed Processing Environment* (TINA DPE) stützt sich beispielsweise auf CORBA (*Common Object Request Broker Architecture*, [OMG97]) von OMG (*Object Management Group*) ab, und die Schnittstellen sogenannter *computational objects* in TINA werden mit Hilfe von TINA ODL (*Object Definition Language*), einer Abwandlung der OMG IDL (*Interface Definition Language*), beschrieben.

Vergleichbar mit dem *Conceptual Model* im IN (INCM) wird bei TINA mit dem *Basic Reference Model* von ODP ein Rahmen für die Spezifikation vorgegeben. Ähnlich wie auf den verschiedenen Ebenen des INCM wird ein System von verschiedenen Standpunkten (*viewpoints*) aus beschrieben. Folgende *viewpoints* sind definiert: *enterprise viewpoint*, *information viewpoint*, *computation viewpoint*, *engineering viewpoint* und *technology viewpoint*. Daneben wird die TINA-Architektur in drei Teilbereiche zerlegt, und zwar bezüglich Verarbeitung (*Computing Architecture*), Diensten (*Service Architecture*) und Netz (*Network Architecture*). Durch die

41. TINA-C Dokumente sind vom WWW-Server www.tinac.com erhältlich.

42. vgl. S. 27

Trennung der verschiedenen Bereiche soll eine modulare und flexible Architektur erzielt werden. Die *Computing Architecture* definiert im wesentlichen die Plattform für die verteilte Verarbeitung (DPE), die *Service Architecture* legt Konzepte und Richtlinien für den gesamten Lebenszyklus von Telekommunikationsdiensten fest, und die *Network Architecture* modelliert die Dienste des Transportnetzes.

Die TINA *Service Architecture* [TINA-SA5] führt den Begriff der Sitzung (*session*) ein. Solch eine Sitzung verallgemeinert und erweitert den traditionellen Begriff des Rufes. Dabei wird außerdem eine Trennung von Dienstzugang (*access*) und Dienstnutzung (*usage*) vorgenommen. Unter Dienstzugang wird das Auffinden und das Anfordern von Diensten verstanden. Bei der Dienstnutzung wird zwischen *service* (Dienst) im Sinne von Zusatzdienst und *communications* (Kommunikation) im Sinne von Kommunikationsdienst differenziert.⁴³ Entsprechend werden Zugangssitzung (*access session*), Dienstsitzung (*service session*) und Kommunikationssitzung (*communication session*) unterschieden. Im Gegensatz zum traditionellen IN, bei dem der Teilnehmer zuerst einen Kommunikationsdienst anfordert und dann gegebenenfalls zusätzliche Dienstlogik aktiviert wird, ist die Abfolge bei TINA gerade umgekehrt. Bevor eine Kommunikationssitzung etabliert werden kann, muß eine Dienstsitzung existieren, welche wiederum auf eine Zugangssitzung angewiesen ist.⁴⁴ Abbildung 2-18 illustriert vereinfachend die verschiedenen Typen von Sitzungen und ihre Beziehungen bei Inanspruchnahme eines Dienstes.

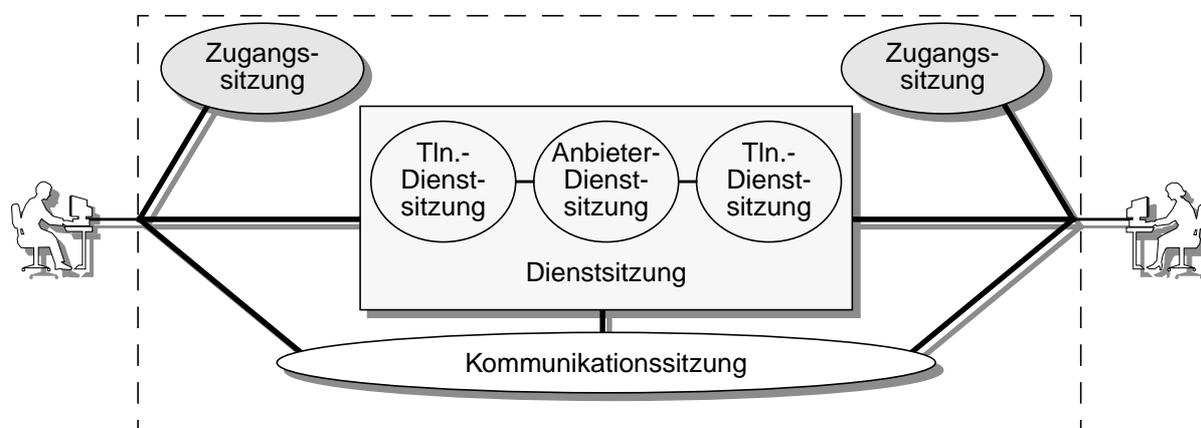


Abbildung 2-18: Vereinfachende Darstellung von TINA-Sitzungen bei Inanspruchnahme eines Dienstes gemäß [DH97, MM96].

Die Betrachtung der *Service Architecture* aus der Sicht des *computational viewpoint* kann am ehesten mit der Betrachtung des IN-Konzepts auf der verteilten funktionalen Ebene verglichen werden. Zur Modellierung werden sogenannte Dienstkomponenten (*service components*) verwendet. Vergleichbar mit funktionalen Instanzen im IN, ermöglichen sie die Kapselung von Daten und Funktionalitäten in einem TINA-basierten System.⁴⁵ Ein Dienst wird durch Interaktion mehrerer Dienstkomponenten erbracht. Die Beschreibung einer Dienstkomponente erfolgt

43. Siehe Diskussion des Dienstbegriffs im IN.

44. Dabei kann der „Besitzer“ einer Dienstsitzung – und damit die zugehörige Zugangssitzung – wechseln.

ebenso wie die der *computational objects*, d. h. mit Hilfe von TINA ODL. Dienstkomponenten werden entsprechend ihrer Zugehörigkeit zu verschiedenen Typen von Sitzungen den Bereichen Zugang (*access*), Nutzung (*usage*) und Kommunikation (*communications*) zugeordnet. Darüberhinaus sind die Dienstkomponenten verschiedenen Bereichen zugeordnet, die den Einflußbereich verschiedener „Rollen“ (*roles*) darstellen. Im folgenden Beispiel seien dies der Teilnehmerbereich (*user domain*) und der Anbieterbereich (*provider domain*).

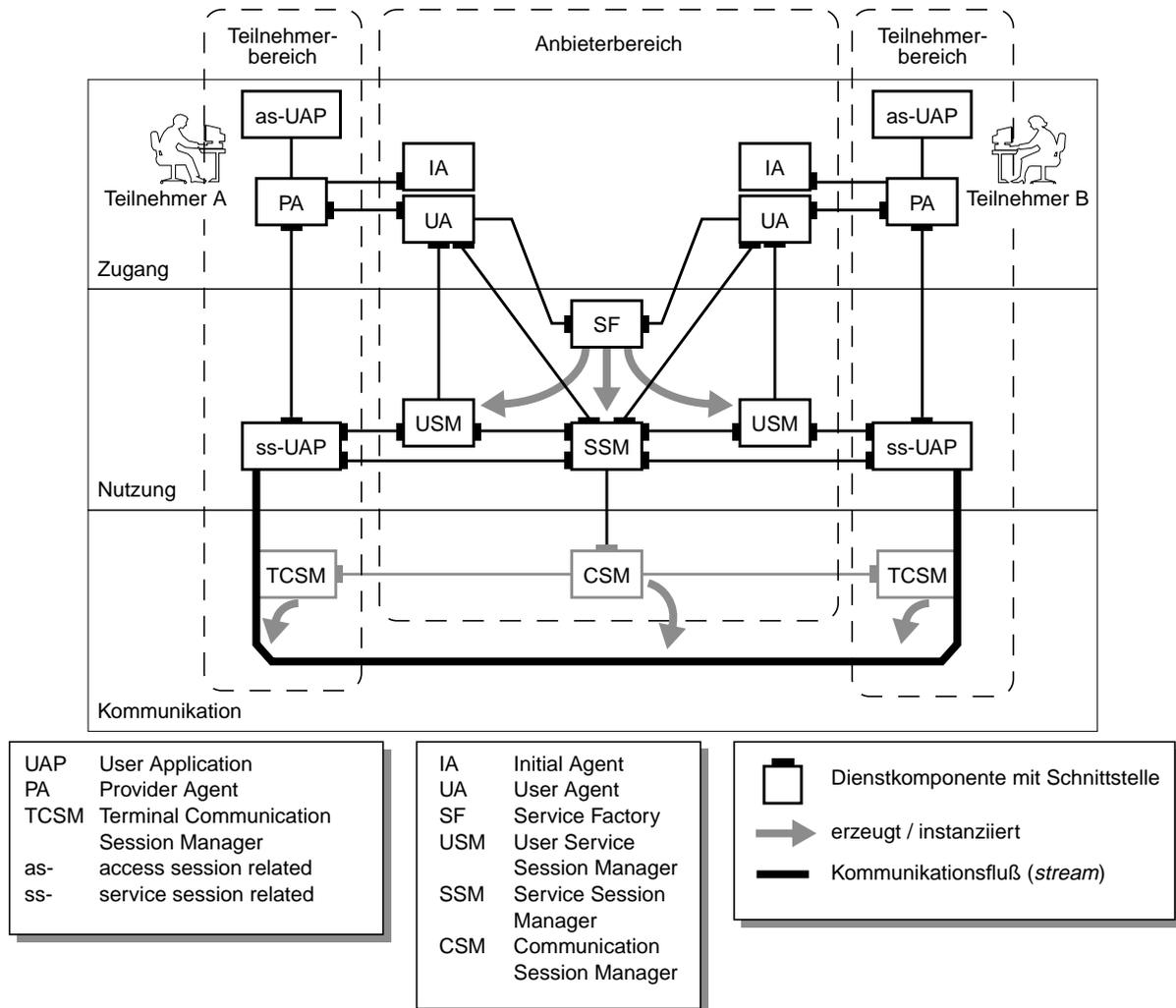


Abbildung 2-19: Prinzipielles Zusammenwirken von TINA-Dienstkomponenten am Beispiel der Inanspruchnahme eines Dienstes eines Anbieters durch zwei Teilnehmer nach [TINA-SA5].

Abbildung 2-19 zeigt anhand eines Beispiels aus [TINA-SA5], wie verschiedene Dienstkomponenten bei der Inanspruchnahme eines Dienstes zusammenwirken. Will sich ein Teilnehmer in ein System einbuchen, so kontaktiert die Anwendung beim Teilnehmer (*access session related User Application*, as-UAP) den Vertreter eines Anbieters (*Provider Agent*, PA), der wie-

45. Eine Dienstkomponente stellt eine höhere Abstraktion als eine funktionale Instanz dar. Eine Dienstkomponente kann auf mehrere *computational objects* bzw. *computational object groups* abgebildet werden, welche selbst wieder Daten und Funktionalität kapseln. Im Gegensatz zu funktionalen Instanzen kann eine Dienstkomponente (im *engineering viewpoint*) auf mehrere Rechner verteilt sein. Unter diesem Aspekt sind *computational objects* mit funktionalen Instanzen vergleichbar, da auch sie nicht aufgeteilt und auf mehrere Rechner verteilt werden können.

derum den Kontakt zu einem Erstkontaktagenten (*Initial Agent*) im Anbieterbereich herstellt. Kann der Teilnehmer sich authentisieren, so ist die Zugangssitzung erfolgreich eröffnet, und es wird ihm ein Vertreter im Anbieterbereich (*User Agent, UA*) zugewiesen. Soll ein Dienst in Anspruch genommen werden, instanziiert der PA eine dienstspezifische Anwendung beim Teilnehmer (*service session related UAP, ss-UAP*) und die zum geforderten Dienst gehörige Diensterzeugungsinstanz (*Service Factory, SF*) wird via den UA kontaktiert, wodurch eine Dienstsitzung eröffnet wird. Die SF instanziiert nun einen Dienstmanager (*Service Session Manager, SSM*) und einen teilnehmerbezogenen Dienstmanager (*User Service Session Manager, USM*), mit denen der ss-UAP sich dann über den zu erbringenden Dienst austauscht. In ähnlicher Weise wird der Teilnehmer B via seinen Vertreter (UA) zu der Dienstsitzung „eingeladen“. Schließlich wird eine Kommunikationssitzung zwischen den beiden Zielsuchern eröffnet indem die beiden ss-UAPs durch einen Kommunikationsfluß (*stream*) verbunden werden. Die die Kommunikation betreffenden Abläufe sind nicht mehr Teil der *Service Architecture*, sondern werden der *Network Architecture* zugeordnet. Aus diesem Grund sind die entsprechenden Dienstkomponenten in Abbildung 2-19 blaß dargestellt.

Im folgenden Unterkapitel wird sowohl für klassische als auch für TINA-basierte „Intelligente Netze“ dargelegt, welche Möglichkeiten und Vorschläge es zur Mobilitätsunterstützung im Rahmen der jeweiligen Konzepte und Standards gibt.

2.4.4 Mobilitätsverwaltung im Rahmen des „Intelligenten Netzes“

Eine Reihe von Projekten und Veröffentlichungen beschäftigen sich mit der Frage, wie Mobilität im Rahmen des „Intelligenten Netzen“ unterstützt werden kann. Eine kleine Auswahl sei hier aufgeführt: [RT96, FBWC97, Gun97, Kle97, Sch96, SB95, AMM96, MN96].

Will man Mobilitätsunterstützung im *Capability Set 1* des Intelligenten Netzes realisieren, so stößt man an gewisse Grenzen. Da keine rufunabhängigen Dienstauslöser möglich sind, ist eine Ortsaktualisierung immer mit einem Rufaufbauversuch verbunden. Die Interaktion mit Endgeräten oder Teilnehmern kann auf Grund fehlender Schnittstellen im wesentlichen nur über die SRF erfolgen. Die wenigen Schnittstellen und insbesondere der fehlende Informationsfluß zwischen funktionalen Einheiten des gleichen Typs erschweren eine effektive verteilte Mobilitätsverwaltung. Will man beispielsweise die Protokollabläufe des MAP für einen mobilen Teilnehmer im Festnetz nachbilden, so wird dies durch eine fehlende Schnittstelle zwischen SCFs beschränkt. Die Steuerfunktionen von VLRs und HLRs müssen dabei von SCFs übernommen werden, die Datenverwaltungsfunktionen von SDFs. Im Gegensatz zu GSM kann es aber je Szenario nur eine steuernde Instanz geben.

Abbildung 2-20 zeigt den Protokollablauf beim Aufenthaltsortwechsel bei IN-basierter Mobilitätsverwaltung analog zur Darstellung in Abbildung 2-8, wiederum unter Verwendung einer temporären Teilnehmererkennung (TMSI). Die Dienststeuerung wird von einer SCF in der Nähe des neuen Aufenthaltsorts ($SCF_{neu}^{visited}$) übernommen. Die Datenverwaltungsfunktionen von VLR_{alt} , VLR_{neu} und HLR entsprechen $SDF_{alt}^{visited}$, $SDF_{neu}^{visited}$ beziehungsweise SDF_{home} .

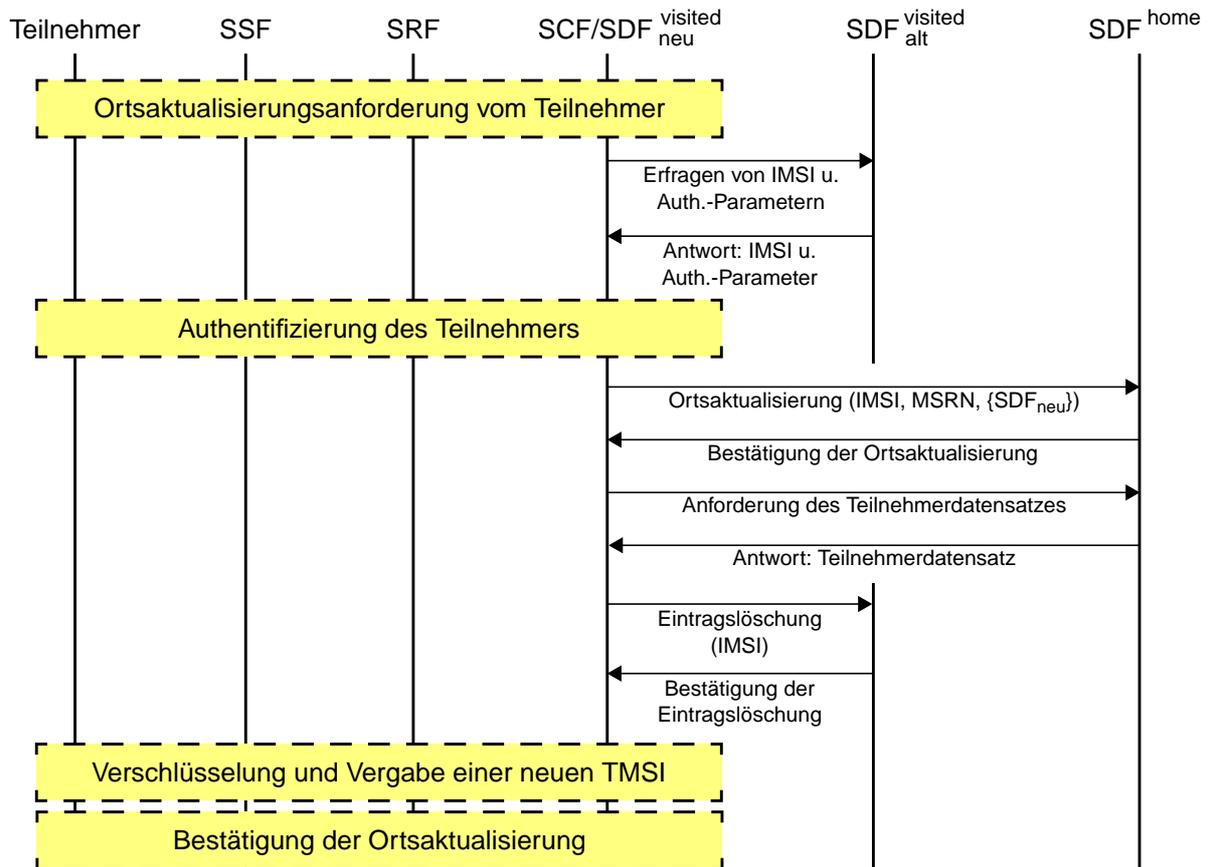


Abbildung 2-20: Protokollablauf beim Aufenthaltsortwechsel bei IN-basierter Mobilitätsverwaltung analog zur Darstellung in Abbildung 2-8.

Alle dargestellten Meldungen beruhen ausschließlich auf Informationsfluß zwischen SCF und SDF (*Add Entry*, *Modify Entry*, *Remove Entry* und *Search*). Es werden also lediglich Operationen auf Datenbankeinträge vorgenommen. Nicht dargestellt ist Informationsfluß zwischen SCF und SDF am neuen Aufenthaltsort sowie zur Kommunikation mit dem Teilnehmer.

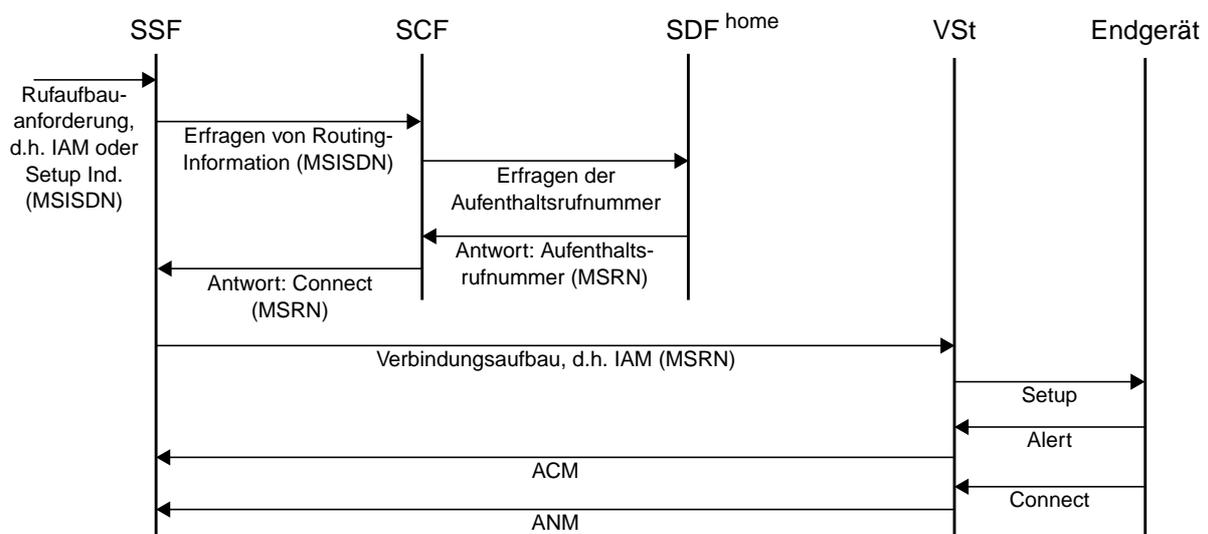


Abbildung 2-21: Protokollablauf beim Aufbau eines Rufes zu einem mobilen Teilnehmer bei IN-basierter Mobilitätsverwaltung analog zur Darstellung in Abbildung 2-9.

Abbildung 2-21 zeigt den Protokollablauf beim Aufbau eines Rufes zu einem mobilen Teilnehmer analog zur Darstellung in Abbildung 2-9. Die Dienststeuerung wird von einer SCF übernommen, die sich entweder nahe beim rufenden Teilnehmer befindet, oder aber, wie bei GSM, mit der Heimatdatenverwaltung (SDF^{home}) gekoppelt ist. Wird in einer SSF eine Rufaufbauanforderung zu einem mobilen Teilnehmer detektiert, so wird die Rufsteuerung an die SCF übergeben (*Initial Detection Point*). Diese verschafft sich nun die Aufenthaltsrufnummer des Teilnehmers von dessen SDF^{home} und weist die SSF an, den Anrufer dorthin zu verbinden (*Connect*). Der weitere Rufaufbau geschieht dann mit Hilfe gewöhnlicher Signalisierung (z.B. ISUP und DSS 1). Hier wurde nun angenommen, daß die Aufenthaltsrufnummer des Teilnehmers dauerhaft vergeben wird, so daß eine Nachfrage am aktuellen Aufenthaltsort nicht notwendig ist.

Im CS-2 wird die Mobilitätsverwaltung besser unterstützt. Das Anzeigen eines Aufenthaltsortswechsels wird durch die Instanzen zur rufunabhängigen Diensterbringung vereinfacht. Der mögliche Informationsfluß zwischen SDFs bzw. SCFs eröffnet eine Vielzahl von Möglichkeiten, um verteilte Mobilitätsverwaltung zu realisieren. SDF-SDF-Informationsfluß dient dazu, die Lokalisation von Daten vor der Dienstlogik zu verbergen und gleichzeitig eine effiziente, verteilte Datenhaltung zu ermöglichen. Dazu sind folgende Informationsflußkomponenten besonders nützlich:

- *Chaining Request*: Kann eine SDF eine Anfrage nicht beantworten, so kann sie diese an eine andere weiterreichen, welche dann wiederum die Antwort auf die Anfrage zurückmeldet (*Chaining Result*)⁴⁶.
- *Copy Request* und *Copy Update*: Eine SDF kann die Kopie eines Datensatzes von einer anderen anfordern (*Copy Request*) und dabei festlegen, unter welchen Bedingungen sie über Änderungen des Datensatzes informiert werden will. Tritt solch eine Änderung ein, wird sie dann von der zuvor befragten SDF informiert (*Copy Update*).

Auch einige Erweiterungen der SCF-SDF-Beziehung unterstützen verteilte Datenhaltung:

- *Add Entry Referral*, *Modify Entry Referral*, *Remove Entry Referral* und *Search Referral*: Die CS-1 Informationsflußkomponenten *Add Entry*, *Modify Entry*, *Remove Entry* und *Search* können damit zurückgewiesen werden und dabei einen Verweis auf eine andere SDF liefern.
- *Execute*: Stößt die Ausführung eines Skripts zum Bearbeiten von Daten in der SDF an. (Dazu ebenfalls: *Execute Referral*).

Mit Hilfe von *Execute* wäre es auch denkbar, ähnlich wie in GSM, Aufenthaltsrufnummern auf Rufbasis zu vergeben, was im CS-1 eine nicht spezifizierte, zusätzliche Funktionalität in einer SDF erfordert hätte, die dies bei einer *Search*-Operation hätte tun müssen.

46. Im folgenden wird die zu jeder Informationsflußkomponente gehörige Ergebnisinformationsflußkomponente (*Result*) nicht mehr aufgeführt.

Desweiteren sieht IN CS-2 Informationsfluß zwischen SCFs vor. Dieser kann beispielsweise dazu verwendet werden, eine entfernte SCF zu beauftragen, auf eine in der Nähe befindliche SDF zuzugreifen und das Ergebnis einer unter Umständen komplexen Interaktion zurückzuliefern. Eine solche SCF wird *unterstützende SCF* genannt. Sie kann entweder sofort antworten oder parallel zur *steuernden SCF* arbeiten und das Ergebnis zu einem späteren Zeitpunkt asynchron melden. Die wichtigsten Informationsflußkomponenten sind:

- ❑ *SCF Bind Request, SCF Unbind Request* und *Activity Test*: Etabliert, beendet bzw. überprüft das Bestehen einer Beziehung zwischen einer steuernden SCF und einer unterstützenden SCF.
- ❑ *Handling Information Request* (inkl. *Handling Information Request Referral*): Stellt die Anfrage der steuernden SCF an die unterstützende SCF.
- ❑ *Request Notification, Notification Provided* und *Confirmed Notification Provided*: Erbittet die unterstützende SCF-Benachrichtigung (*Request Notification*), so erfolgt bei entsprechenden Ereignissen eine solche (*Notification Provided* bzw. *Confirmed Notification Provided*); wobei eine Benachrichtigung u. U. bestätigt werden muß.
- ❑ *Provide User Information* und *Additional Information Result*: Damit fordert die unterstützende SCF weitere Informationen über einen Teilnehmer von der steuernden SCF an, welche sie dann als *Additional Information Result* erhält.
- ❑ *Network Capability Request*: Damit fordert die unterstützende SCF-Informationen über die von der steuernden SCF erbringbaren Dienste an.

Eine IAF kann auch die Rolle einer unterstützenden SCF übernehmen und damit zum Beispiel einen Übergang zwischen einer IN-basierten Mobilitätsverwaltung und einem GSM-Netz eröffnen. Zusammenfassend kann man also sagen, daß die Empfehlungen des IN CS-2 voraussichtlich eine hervorragende Ausgangsbasis für verteilte Mobilitätsverwaltung und für Interworking mit anderen Netzen bieten werden.

Auch für TINA ist eine spezielle Mobilitätsverwaltung notwendig. Soll beispielsweise der *User Agent* eines Teilnehmers von einem *Service Session Manager* „eingeladen“ werden, so muß er adressiert werden können. Dazu muß dem SSM seine Adresse vom *User Agent* des einladenden Teilnehmers übergeben werden; dieser wiederum muß sich zuvor die entsprechende Adresse beschaffen. In [TINA-SA5] heißt es dazu in einer Fußnote: „*The UA may use a location service for locating user B's UA, or some other means*“. Existiert noch keine Zugangssitzung für den gerufenen Teilnehmer, so wird u. U. ein Endgerät, an dem dieser Teilnehmer in irgendeiner Weise registriert ist, alarmiert. Dieses kann den Teilnehmer dann auffordern, eine Zugangssitzung zu initiieren. Auch hier muß die Adresse und der Ort dieses Endgerätes ermittelt werden können. In dem *TINA-C auxiliary project „Personal Communication Support in TINA“* werden diese Fragen näher untersucht und beispielsweise ein sogenannter *PCS-enhanced User Agent* (PCS-UA) definiert, der persönliche Mobilität entsprechend unterstützt [EM96, EMPZ96].

2.5 Mobilität in IP-basierten Netzen: *Mobile IP*

Standardisierung für das Internet erfolgt im allgemeinen durch die *Internet Engineering Task Force* (IETF). In verschiedenen Arbeitsgruppen werden Vorschläge für neue Internet-Normen erarbeitet. Diese Vorschläge durchlaufen mehrere Stufen eines Standardisierungsprozesses: vom sogenannten *Internet-Draft* über den *Request for Comments* (RFC), den *Proposed Standard* und den *Draft Standard* zum *Internet Standard*. Aus der Arbeitsgruppe „*IP Routing for Wireless/Mobile Hosts (mobileip)*“ ist der RFC 2002 „*IP Mobility Support*“ [Per96],⁴⁷ der bisher den Status eines *Proposed Standard* erreicht hat, hervorgegangen. Man bezeichnet das zugehörige Protokoll üblicherweise als *Mobile IP*. Das Ziel von *Mobile IP* ist es, Rechnern unter Beibehaltung ihrer Netzadresse⁴⁸ den Wechsel des Netzzugangspunktes – und damit Mobilität – zu ermöglichen. Eine Übersicht zu *Mobile IP* findet sich in [Per97].

Mobile IP ist ein Beispiel dafür, wie Mobilitätsunterstützung außerhalb beziehungsweise am Rand eines Netzes erfolgt. *Mobile IP* muß lediglich im mobilen Rechner und in seinem Stellvertreter im Heimatnetz⁴⁹, dem *Home Agent*, unterstützt werden. Ansonsten bleiben die Internet-Protokolle in anderen Rechnern bzw. Routern unverändert. Häufig befindet sich darüberhinaus im besuchten Netz eine weitere Instanz, die *Mobile IP* unterstützt, der sogenannte *Foreign Agent*, welcher die Adreßvergabe vereinfacht und für den mobilen Rechner zumeist die Rolle des nächstgelegenen Routers übernimmt.⁵⁰

2.5.1 Die Funktionsweise von *Mobile IP*

Die prinzipielle Funktionsweise von *Mobile IP* ist relativ einfach. Verschiebt ein mobiler Rechner IP-Pakete, so wird unabhängig von seinem Netzzugangspunkt seine permanente IP-Adresse als Absenderadresse verwendet. Pakete, die für einen mobilen Rechner bestimmt sind, tragen diese permanente IP-Adresse als Zieladresse und werden dementsprechend ins Heimatnetz des mobilen Rechners geführt. Befindet der mobile Rechner sich nicht in seinem Heimatnetz, so empfängt dort der *Home Agent* die IP-Pakete an seiner Stelle. Im besuchten Netz verschafft sich der mobile Rechner entweder mit Hilfe von Protokollen, die nicht Teil von *Mobile IP* sind, eine lokale IP-Adresse, oder er ermittelt die IP-Adresse eines *Foreign Agent* im besuchten Netz, der bereit ist, Pakete für ihn entgegenzunehmen. In beiden Fällen wird die gewonnene IP-Adresse dem *Home Agent* als sogenannte *care-of address* mitgeteilt. Diesem wird dadurch möglich, die für den mobilen Rechner bestimmten IP-Pakete an die *care-of*

47. Zugehörige sind RFC 2003, RFC 2004, RFC 2005 und RFC 2006.

48. Der Adresse des Dienstzugangspunktes der Vermittlungsschicht; hier: der IP-Adresse.

49. Das Internet wird durch eine Vielzahl von sogenannten „Netzen“ gebildet. Eine IP-Adresse untergliedert sich in *netid* und *hostid*. (Welcher Teil der Adresse als *netid* und welcher als *hostid* interpretiert wird, hängt von der Adreßklasse und der sogenannten Subnetzmaske ab). Verkehrslenkung erfolgt auf Basis der *netid* bis zu einem Router, der sich im selben „Netz“ (*net*) wie der adressierte Rechner (*host*) befindet, d.h. der direkt mit dem adressierten Rechner über ein Protokoll der Schicht 2 kommunizieren kann. Alternativ wird das Internet auch als ein Netz bezeichnet, das von mehreren „Subnetzen“ gebildet wird.

50. Zumeist ist die *Foreign-Agent*-Funktionalität in einem Router des besuchten Netzes lokalisiert. Im folgenden wird der *Foreign Agent* jedoch verallgemeinernd als unabhängig von einem Router dargestellt.

address weiterzuleiten. Dazu bedient der *Home Agent* sich der Technik der Kapselung. Bei IP-in-IP-Kapselung werden Paketkopf und Paketinhalt eines IP-Pakets als Paketinhalt in ein neues IP-Paket verpackt, das dann einen eigenen, neuen Paketkopf erhält. Bei *Mobile IP* wird das für den mobilen Rechner bestimmte Paket in ein Paket verpackt, das als Absenderadresse die IP-Adresse des *Home Agent* und als Zieladresse die *care-of address* trägt. Das ursprüngliche Paket wird am Ziel, d.h. entweder beim mobilen Rechner oder beim *Foreign Agent*, wieder ausgepackt und über den ursprünglich adressierten IP-Dienstzugangspunkt ausgeliefert, wobei im Falle eines beteiligten *Foreign Agent* das Paket zuerst über die Schicht 2 zum mobilen Rechner weitergereicht wird. Verallgemeinernd nennt man diese Vorgänge auch „Tunneln“. „Tunneln“ ist deshalb erforderlich, weil für die IP-Dienstbenutzer *Mobile IP* nicht sichtbar sein darf. Dazu gehört auch, daß der IP-Paketkopf nicht verändert werden darf. So benutzt beispielsweise das Internet-Transportprotokoll TCP (*Transmission Control Protocol*) die IP-Absenderadresse und die IP-Zieladresse um eine TCP-Verbindung zu identifizieren. Darüberhinaus werden diese Adressen von TCP auch bei der Bildung einer Prüfsumme mitverwendet.

Abbildung 2-22 zeigt die Wegführung von IP-Paketen in *Mobile IP* bei bidirektionaler Kommunikation zwischen einem nicht-mobilen Rechner und einem mobilen Rechner, der im besuchten Netz von einem *Foreign Agent* unterstützt wird. Während die Pakete vom mobilen

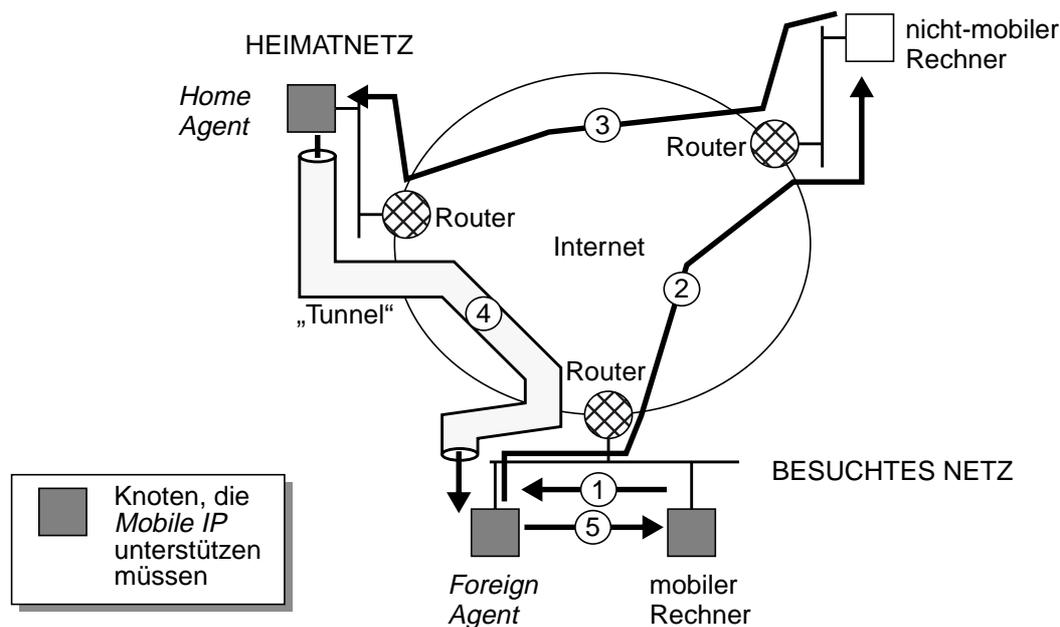


Abbildung 2-22: Beispiel für die Wegführung von IP-Paketen in *Mobile IP* bei bidirektionaler Kommunikation zwischen einem nicht-mobilen Rechner und einem mobilen Rechner, der in einem besuchten Netz von einem *Foreign Agent* unterstützt wird.

Rechner über den *Foreign Agent* (①), der dem mobilen Rechner als nächster Router bekannt ist, auf direktem Weg (②) zum Ziel geführt werden können, nehmen Pakete zum mobilen Rechner einen Umweg über den *Home Agent* (③), der sie dann zum *Foreign Agent* „tunnelt“ (④), von wo sie im lokalen Netz über die Schicht 2 zum mobilen Rechner gelangen (⑤). Man spricht in diesem Zusammenhang auch von „Dreiecks-Routing“. *Mobile IP* kann mit einer Rufweiterleitung im ISDN verglichen werden. Diese Analogie läßt sich sowohl für das ver-

mittlungsstellengestützte Rufweiterleiten als auch für das Weiterleiten in einem ISDN-Endgerät ziehen, je nachdem, ob die Funktionalität des *Home Agent* in einem Router oder in einem gewöhnlichen Rechner lokalisiert ist.

Eine detailliertere Betrachtung der Funktionsweise von *Mobile IP* erfordert spezifischere Kenntnisse der TCP/IP-Protokolle. Dazu wird auf [Com91] verwiesen. Ein wichtiger Mechanismus bei *Mobile IP* ist das sogenannte *Agent Advertisement*. Ein mobiler Rechner muß erkennen können, ob er sich in ein neues Netz begeben hat. Mit Hilfe von erweiterten Meldungen des *Internet Control Message Protocol* (ICMP) geben *Home Agents* und *Foreign Agents* periodisch ihr Vorhandensein und ihre IP-Adresse bekannt. Dadurch kann der mobile Rechner solch einen Netzwechsel erkennen. Unter Verwendung von *ICMP Router Solicitation* Meldungen kann ein mobiler Rechner diese Agenten auch zum Aussenden der Informationen auffordern. Dem *Home Agent* muß dann die neue *care-of address* mitgeteilt werden; es muß eine „Registrierung“ erfolgen. Dies geschieht üblicherweise via den *Foreign Agent* mit Hilfe von UDP-Paketen (*User Datagram Protocol*) an Port 434. Dabei wird auch die Lebensdauer der *care-of address* ausgehandelt. Timer im mobilen Rechner, im *Foreign Agent* und im *Home Agent* überwachen die Lebensdauer. Vor Ablauf im *Home Agent* wird vom mobilen Rechner erneut eine Registrierung durchgeführt. „Deregistrierung“ beim *Home Agent* wird durchgeführt, wenn der mobile Rechner in sein Heimatnetz zurückkehrt. *Foreign Agents* werden über Ortswechsel nicht informiert. Dort erfolgt die Deregistrierung automatisch nach Ablauf der Lebensdauer. Registrierung und Deregistrierung dürfen nicht von anderen Rechnern zum Umleiten von IP-Paketen mißbraucht werden können. Dazu erfolgt eine Sicherung der Authentizität von Nachrichten. Die verwendeten Verfahren beruhen zumeist auf einem gemeinsamen Geheimnis, das im voraus außerhalb des Internet ausgetauscht wird. Zum Schutz gegen Wiederverwendung von Nachrichten werden Zeitstempel oder auch sogenannte *Nonces* verwendet. *Nonces* sind Zufallszahlen, die eine Instanz der anderen zuschickt und die diese in der darauffolgenden Nachricht wieder zurückschicken muß.⁵¹ Innerhalb eines fremden Netzes darf ein mobiler Rechner nicht am *Address Resolution Protocol* (ARP) teilnehmen, da er sonst von anderen Rechnern direkt im lokalen Netz adressiert würde und dies auch noch nach dem Verlassen des besuchten Netzes so erfolgen würde. Im Heimatnetz muß beim Verlassen des mobilen Rechners mit Hilfe von *gratuitous ARP*-Nachrichten dafür gesorgt werden, daß der *Home Agent* anstelle des mobilen Rechners direkt adressiert wird. Bei der Rückkehr des mobilen Rechners wird in gleicher Weise dafür gesorgt, daß dieser wieder direkt adressiert wird.

Neben der nicht optimalen Verkehrsführung (Dreiecks-Routing) hat *Mobile IP* noch weitere unschöne Eigenschaften. Dadurch, daß der mobile Rechner Pakete an seine Kommunikationspartner unter Verwendung seiner permanenten IP-Adresse als Absenderadresse schickt, werden Pakete erzeugt, die aus Netzen kommen, aus denen sie gemäß ihrer Absenderadresse gar nicht kommen dürften, und die u. U. von außen in ein Netz geschickt werden, aus dem sie

51. Schickt ein Mobile Node eine Registrierungs-Meldung ohne einen gültigen *Nonce* an den *Home Agent*, so antwortet dieser mit einem *Registration Reject*, das einen neuen *Nonce* enthält, mit dem dann eine erfolgreiche Registrierung durchgeführt werden kann.

gemäß ihrer Absenderadresse eigentlich stammen sollten. In beiden Fällen kann es deshalb vorkommen, daß Router oder Firewall-Rechner diese IP-Pakete verwerfen, so daß keine Kommunikation mehr erfolgen kann. Als Abhilfe wird das „Tunneln“ von Paketen vom *Foreign Agent* zum *Home Agent* in Erwägung gezogen. Solch eine Lösung würde jedoch auch für diese Richtung eine nicht-optimale Wegführung bedeuten. Abbildung 2-23 zeigt dieses doppelte „Tunneln“ analog zum Beispiel in Abbildung 2-22.

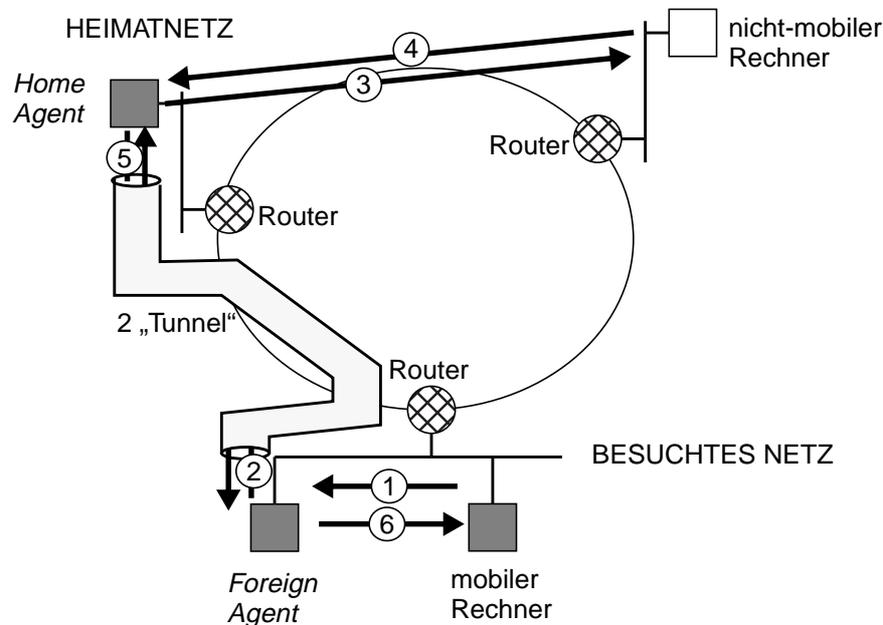


Abbildung 2-23: Beispiel analog zu Abbildung 2-22 für die Wegführung von IP-Paketen in *Mobile IP* bei der Verwendung von doppeltem „Tunneln“.

Verläßt ein mobiler Rechner ein besuchtes Netz, so gehen alle Meldungen, die danach den bisherigen *Foreign Agent* erreichen, verloren. Dies können sehr viele Pakete sein, da beim *Home Agent* erst nach einer erfolgreichen Registrierung im neuen Netz die Pakete dorthin umgeleitet werden. Diese Verluste müssen von den IP-Dienstbenutzern ausgeglichen werden. Bei TCP beeinträchtigen sie die Flußsteuerung und führen kurzzeitig zu einer Verringerung der Senderate bzw. des Sendefensters.⁵²

2.5.2 Die weitere Entwicklung von *Mobile IP*

Für die weitere Entwicklung von *Mobile IP* sind zwei Aspekte von besonderer Bedeutung. Zum einen sollte die Verkehrsführung möglichst optimal erfolgen (*Route Optimization*), d.h. Pakete zu einem mobilen Rechner sollten auf direktem Weg zu seiner momentanen *care-of address* geschickt werden können. Dazu muß jedoch auch der Kommunikationspartner *Mobile IP* unterstützen und in der Lage sein, entsprechend gesicherte Nachrichten mit dem *Home Agent* auszutauschen. Zum anderen sollte ein sanfter Wechsel zwischen Netzen (*smooth han-*

52. Es findet ein sogenannter *slow start* statt. Bei schlechten TCP/IP-Implementierungen können solche Verluste auch zur Blockierung von TCP-Verbindungen führen.

doff) möglich sein. Dazu sollte ein ehemaliger *Foreign Agent* Pakete zur neuen weiterleiten, bis der Ortswechsel beim *Home Agent* bzw. bei anderen Kommunikationspartnern bekannt ist. Auch hier spielt die Sicherung der Authentizität von Nachrichten eine wichtige Rolle.

Beide Aspekte werden in dem *Internet-Draft* „*Route Optimization in Mobile IP*“ [PJ97] behandelt. Eine detaillierte Beschreibung dazu findet sich in Kapitel 3.5.4. der vorliegenden Arbeit.

Kapitel 3

Mobilitätsverwaltung

In diesem Kapitel werden die grundlegenden Aufgaben der Mobilitätsverwaltung in einem Kommunikationsnetz vorgestellt (Unterkapitel 3.1) und es wird gezeigt, daß beim Entwurf eines Verfahrens zur Mobilitätsverwaltung ein angemessenes Gleichgewicht zwischen Aufwand zur Ortsaktualisierung, Aufwand zur Ermittlung der Zieladresse einer mobilen Instanz und Effizienz der Nutzverkehrsführung gefunden werden muß. Unterschiedliche Verfahren lassen sich danach klassifizieren, in welcher Weise das Zusammenspiel von Wegführung und Zielsuche (Unterkapitel 3.2) erfolgt und wie die Datenhaltung (Unterkapitel 3.3) organisiert wird. Die Art der Datenhaltung wird nach drei Kriterien klassifiziert:

- ❑ Verfahren zur Platzierung von Daten (Unterkapitel 3.3.1)
- ❑ Verfahren zum Ausfindigmachen von Daten (Unterkapitel 3.3.2) und
- ❑ Verfahren zur Datenaktualisierung (Unterkapitel 3.3.3).

Der Frage der Rekonfigurierung bestehender Verbindungen bei Netzzugangspunktwechsel, im folgenden als Verbindungsweitergabe (*handover*) bezeichnet, ist das kurze Unterkapitel 3.4 gewidmet. Abschließend werden in Unterkapitel 3.5 Beispiele aktueller oder vorgeschlagener Systeme präsentiert.

3.1 Aufgaben der Mobilitätsverwaltung

Mobilitätsverwaltung in einem Kommunikationsnetz hat die Aufgabe, unabhängig vom Netzzugangspunkt

- ❑ mobile Instanzen erreichbar zu machen („kommende Rufe“),
- ❑ mobilen Instanzen den Zugang zu Telekommunikationsdiensten zu ermöglichen („gehende Rufe“)
- ❑ Netzzugangspunktwechsel von mobilen Instanzen zu erkennen, zu unterstützen und beim Wechsel bestehende Verbindungen aufrecht zu erhalten (Verbindungsweitergabe, „*handover*“).

Die vorliegende Arbeit beschäftigt sich im wesentlichen mit Verfahren, die dazu dienen, mobile Instanzen erreichbar zu machen. Dazu gehört insbesondere das Ermitteln der Adresse, unter der eine mobile Instanz zu einem bestimmten Zeitpunkt erreichbar ist. Diese Aufgabe der

Mobilitätsverwaltung soll im folgenden – analog zur Teilnehmerlokalisierung in GSM – als *Ziellokalisierung* bezeichnet werden. Die *Ziellokalisierung* läßt sich in *Ortsaktualisierung* und *Zielsuche* untergliedern. Die Mobilitätsverwaltung trägt außerdem dazu bei, daß Daten, die benötigt werden, um einer mobilen Instanz Zugang zu Diensten zu gewähren, an ihrem momentanen Netzzugangspunkt verfügbar sind. Dieser Aspekt wird in den folgenden Unterkapiteln am Rande mitbehandelt.

Die Diskussion von Mobilität im OSI-Referenzmodell (Kapitel 2.1.1) hat gezeigt, daß wesentliche Teile der Mobilitätsverwaltung in Kommunikationsnetzen am besten in der Vermittlungsschicht erfolgt. Dabei erzwingen hierarchische Adressierung und Wegführung sowie die Verwendung permanenter Schicht-3-Adressen eine Trennung von Dienstzugangspunktadresse und Wegführungsinformation. In bestehenden Systemen wird die Dienstzugangspunktadresse häufig solange zur Verkehrslenkung verwendet, bis eine Instanz in der Lage ist, diese Adresse in eine zur Zielführung geeignete Adresse umzusetzen. Auf welche Adresse diese Umsetzung erfolgt, hängt vom momentanen Netzzugangspunkt der mobilen Instanz ab. Dazu werden Daten über den momentanen Netzzugangspunkt im Netz gehalten oder bei Bedarf ermittelt. Für Ort und Art der Datenhaltung gibt es eine Vielzahl von Möglichkeiten, welche den Aufwand und die Protokolle zum Ausfindigmachen und Aktualisieren dieser Daten stark beeinflussen. Eine leistungsfähige Mobilitätsverwaltung muß dabei ein angemessenes Gleichgewicht zwischen Aufwand zur Ortsaktualisierung, Aufwand zur Ermittlung der Zieladresse und Effizienz der Nutzverkehrsführung finden.

Abbildung 3-1 illustriert vier verschiedene Realisierungsformen in diesem Spannungsfeld. In den Fällen (a) und (b) wird die Ortsinformation an einem zentralen Ort gehalten. Der Aufwand zur Ortsaktualisierung hängt damit von der Entfernung der mobilen Instanz vom Ort der Datenhaltung ab. Im Fall (a) wird der Nutzverkehr in die Nähe der Ortsinformation geführt, was eine erhöhte Belegung von Ressourcen für den Nutzverkehr verursachen kann, jedoch für die eigentliche Zielsuche nur geringen Aufwand entstehen läßt. Umgekehrt wird im Fall (b) die Zielsuche vom Ursprung des Nutzverkehrs aus gestellt, was den Aufwand für dieselbe erhöht, jedoch eine direkte Nutzverkehrsführung ermöglicht. Auch in den Fällen (c) und (d) erfolgt direkte Nutzverkehrsführung. Im Fall (c) wird der Aufwand für die Zielsuche durch Plazierung von Kopien der Ortsinformation in der Nähe potentieller Anfrager verkleinert. Dies vergrößert jedoch entsprechend den Aufwand zur Ortsaktualisierung. Im Fall (d) wird die Ortsinformation näher am aktuellen Aufenthaltsort der mobilen Instanz gehalten. An zentraler Stelle dagegen befindet sich ein Verweiszeiger auf den aktuellen Ort der Datenhaltung, der sich seltener ändert. Hiermit kann der Aufwand zur Ortsaktualisierung verringert werden. Gleichzeitig wird jedoch der Aufwand für die Zielsuche erhöht.

In den nachfolgenden Unterkapiteln werden das Zusammenspiel von Wegführung und Zielsuche sowie verschiedene Verfahren zur Plazierung von Daten, zum Ausfindigmachen von Daten und zur Aktualisierung von Daten vorgestellt.

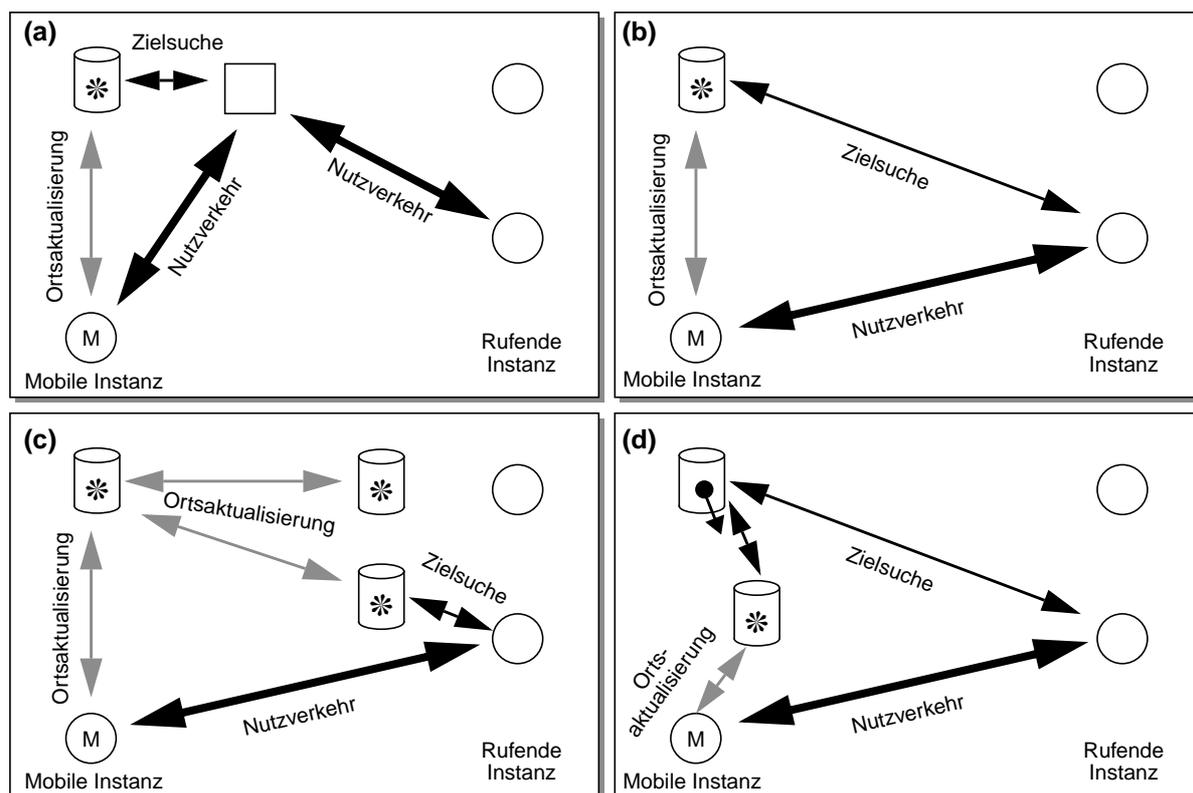


Abbildung 3-1: Unterschiedliche Realisierungsformen – (a) bis (d) – der Ziellokalisierung und ihr Zusammenspiel mit der Nutzverkehrsführung.

3.2 Zusammenspiel von Wegführung und Zielsuche

Klassische Telekommunikationsnetze verwenden zum Austausch von Steuermeldungen im allgemeinen auf das Signalisiersystem Nummer 7. Im Internet stützen sich Steuerungsprotokolle, wie zum Beispiel ICMP, meist direkt auf das Internet Protokoll (IP) ab. In beiden Fällen gilt, daß üblicherweise der Anteil der Steuermeldungen am Gesamtverkehr relativ klein ist. Es erscheint deshalb sinnvoll, zu versuchen, auf Kosten der Steuermeldungen den Nutzverkehr möglichst direkt von seinem Ursprung zu seinem Ziel zu führen. Dazu sollte die Umsetzung auf eine zur Wegführung geeignete Adresse möglichst nahe am Ursprung des Nutzverkehrs erfolgen und die Zielsuche – also die Ermittlung benötigter Ortsinformation – vorab mit Hilfe von Steuermeldungen erfolgen. In realen Systemen findet dies jedoch nicht immer in so idealer Form statt. Dafür gibt es unterschiedliche Gründe.

Ein wichtiger Grund ist, daß oftmals die Mobilität einer Instanz nicht erkannt wird, beziehungsweise keine Unterscheidung von nicht-mobilen Instanzen möglich ist. Dies kann daran liegen, daß Adressen für mobile und nicht-mobile Instanzen aus dem selben Adreßraum vergeben werden. In bestehende Kommunikationsnetze wurde Mobilität zumeist erst nachträglich eingebracht. In Teilen dieser Netze wird sie deshalb häufig nicht unterstützt, so daß Verkehrsführung zuerst auf der Basis permanenter Adressen erfolgt, bis eine mobilitätsbewußte Instanz in der Lage ist, eine Adreßumsetzung durchzuführen. Als Beispiel sei das weltweite Telefonnetz genannt, in dem GSM-Endgeräte von überall erreichbar sind. Ein Ruf wird dabei zuerst

anhand der Teilnehmerrufnummer (MSISDN) in das Heimatnetz des Angerufenen geführt, von wo eine Weiterleitung auf Basis der Aufenthaltsrufnummer (MSRN) in das besuchte Netz erfolgt.

Auch aus Sicherheitsgründen erfolgt oftmals nicht-optimale Wegführung. Kann eine Instanz etwa die Authentizität von Ortsinformation nicht überprüfen, so ist es sinnvoll, Wegführung solange auf Basis der permanenter Adresse durchzuführen, bis zuverlässige Ortsinformation erlangt werden kann. Dies ist der Hauptgrund, warum *Mobile IP* nach RFC 2002 IP-Pakete prinzipiell über das Heimatnetz beziehungsweise über den *Home Agent* führt. Ein ähnlicher Fall liegt vor, wenn die Information über den aktuellen Aufenthaltsort eines Teilnehmers einen bestimmten Bereich nicht verlassen soll. Es liegt nahe, Nutzverkehr dann zuerst in diesen Bereich zu führen und von dort weiterzuleiten. Sollen in einem Kommunikationsnetz Ursprung und Ziel einer Kommunikationsbeziehung verschleiert werden, so können dazu jedoch auch sogenannte MIXe¹ [Cha81, PPW91, PPW89] eingesetzt werden. Ein Beobachter, der keinen Einblick in die Vorgänge innerhalb eines MIXes hat, kann zwischen Verkehr in einen MIX hinein und Verkehr aus einem MIX heraus keinen Zusammenhang herstellen. Wegführungsinformation am Ursprung eines Rufs reicht lediglich bis zum ersten MIX in einer Kette von MIXen („MIX-Kaskade“). Soll Ortsinformation über einen mobilen Teilnehmer wegen ihres vertraulichen Charakters nicht jedem Netzknoten/Endsystem zur Verfügung gestellt werden, so könnten die Mobilitätsverwaltungsprotokolle Rufe über verschiedene MIXe führen, was eine Wegführung ohne Preisgabe von Ortsinformation ermöglichen würde, da die Adresse, die zur Wegführung verwendet wird, die eigentliche Ortsinformation vor dem Absender verbirgt. Weitergehende Konzepte zum Schutz vertraulicher Informationen in Mobilkommunikationssystemen finden sich in [Pfi93, FJKP95, KFJP].

Wird eine Kommunikationsbeziehung auch dem Angerufenen in Rechnung gestellt, so fällt es dem Vertragspartner des Angerufenen – also dem Betreiber seines Heimatnetzes – leichter, den Nachweis über erfolgte Kommunikation zu erbringen, wenn an dieser auch das Heimatnetz beteiligt ist. Ein weiterer Grund, Nutzverkehr über das Heimatnetz des Gerufenen zu führen, liegt vor, wenn es nur dort möglich ist, bestimmte Zusatzdienste zu erbringen. Kann man dagegen Dienste vom Heimatnetz aus steuern, entfällt dieser Grund gegebenenfalls. So ist beispielsweise in der GSM-Phase 2+ mit CAMEL ein Konzept vorgesehen, bei dem IN-Dienste durch das HLR im Heimatnetz gesteuert werden. Sollte jedoch die Verzögerung bei der Signalisierung zwischen HLR und besuchter MSC zur effektiven Dienststeuerung zu groß sein, müßte nach wie vor eine Führung von Rufen über das Heimatnetz erfolgen.

1. MIX ist ein Kunstwort aus dem englischen Verb *to mix*, das in Regel nicht als Akronym angesehen wird.

Zusammenfassend läßt sich sagen, daß optimale Wegführung oftmals zusätzliche Mechanismen im Netz erfordert und unerwünschte Konsequenzen wie das Weitergeben vertraulicher Informationen nach sich ziehen kann. Weicht die Wegführung ohne diese zusätzlichen Mechanismen nur wenig oder nur selten von der optimalen Wegführung ab, so kann der Verzicht auf dieselben gerechtfertigt sein.

In diesem Zusammenhang stellt sich noch die Frage, ob die Zielsuche vor oder gemeinsam mit der Wegführung stattfindet. Erfolgen Zielsuche und Wegführung gemeinsam, so kommt es in der Regel während des Verbindungsaufbaus durch Datenbankabfragen zur Zielsuche zu einer größeren Belegungsdauer von Ressourcen, die beim Verbindungsaufbau reserviert werden. Andererseits kann man bei verbindungsorientierter Kommunikation mit einem geringeren „Rufverzug“ (*post-dialling delay*) rechnen. Im allgemeinen ist jedoch eine klare Trennung zwischen Mobilitätsverwaltung und Verkehrsführung wünschenswert, um die Verkehrsführung unabhängig und flexibel durchführen zu können. Dies wird dadurch erreicht, daß die Zielsuche vor der Wegführung erfolgt, was bei verbindungsorientierter Kommunikation zwar den Rufverzug erhöht, andererseits aber auch die Dauer der Ressourcenbelegung verringert. Abbildung 3-2 veranschaulicht in Signal-Zeit-Diagrammen die beiden beschriebenen Alternativen.

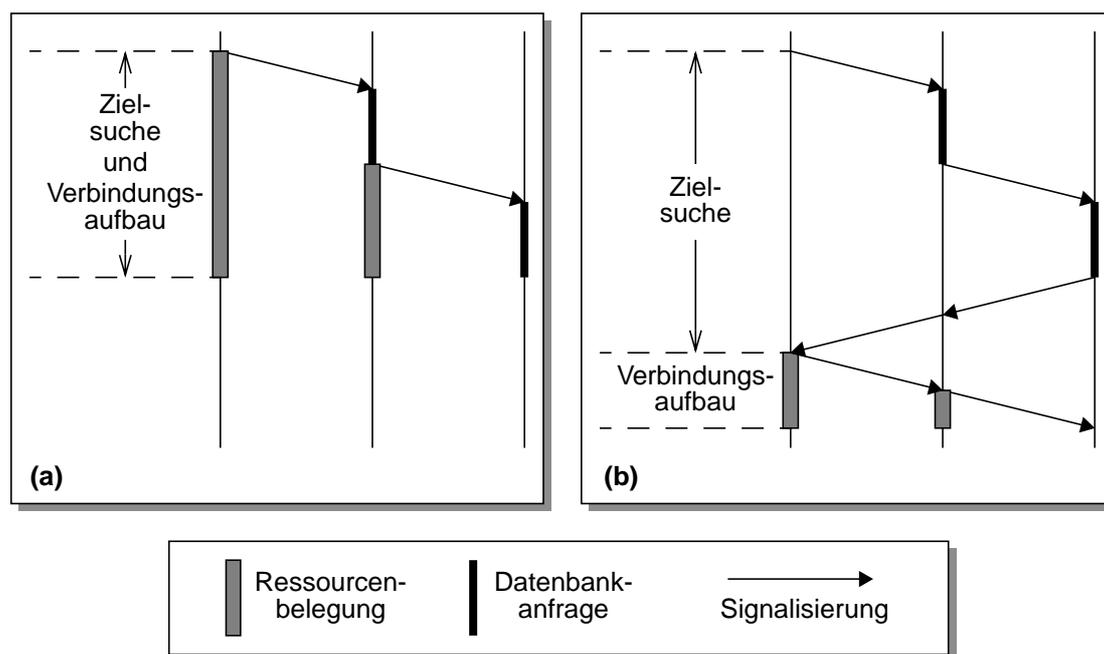


Abbildung 3-2: Veranschaulichung des Zusammenspiels von Zielsuche und Verbindungsaufbau:

- a) Zielsuche und Verbindungsaufbau erfolgen gemeinsam und
- b) Zielsuche erfolgt vor dem Verbindungsaufbau.

Bei verbindungsloser Kommunikation ist der Aufwand zur Beschaffung von Ortsinformation relativ zur Größe eines Datenpaketes unter Umständen unverhältnismäßig groß. Geht man jedoch davon aus, daß im allgemeinen in kurzer Zeit mehrere Pakete an ein bestimmtes Ziel geschickt werden, kann einmal erlangte Ortsinformation wiederverwendet werden, so daß sich auch hier eine Entkopplung von Zielsuche und Wegführung anbietet.²

Erfolgen Zielsuche und Verbindungsaufbau gemeinsam, so muß ein Bezeichner für die Zielinstanz beim Verbindungsaufbau mitgeführt werden. Bei Trennung von Zielsuche und Verbindungsaufbau bieten sich zwei Alternativen. Entweder wird ein entsprechender Bezeichner etwa in Form der permanenten, ortsunabhängigen Adresse mitgeführt (Abbildung 3-3a), oder aber die temporäre, zur Verkehrsführung verwendete Adresse läßt sich am Ziel eindeutig auf einen solchen Bezeichner abbilden (Abbildung 3-3b). Das Mitführen eines Bezeichners hat den Vorteil, daß für mehrere Zielinstanzen dieselbe temporäre Adresse zur Wegführung verwendet werden kann. In manchen Fällen ermöglichen jedoch die verwendeten Protokolle das Mitführen eines solchen Bezeichners nicht. Dann muß die Zuordnung von temporärer Adresse zu einer Zielinstanz entweder wie bei GSM dynamisch je Kommunikationsbeziehung oder aber semi-permanent – d. h. solange die temporäre Adresse für die Verkehrsführung geeignet ist – erfolgen. Das Fehlen eines Bezeichners für die Zielinstanz hat darüberhinaus den Nachteil, daß das Wiederverwenden von temporären Adressen (*Caching*) ohne Konsistenzsicherung nicht mehr sinnvoll möglich ist, da damit unter Umständen eine Kommunikationsbeziehung zu einer „falschen“ Instanz aufgenommen wird.

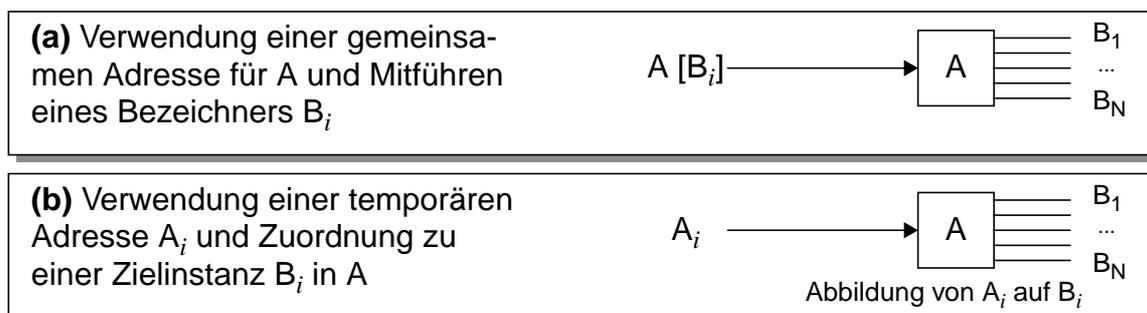


Abbildung 3-3: Alternativen zur Adressierung verschiedener mobiler Instanzen B_i , die über die selbe Zielinstanz (A) erreichbar sind.

3.3 Datenhaltung

Daten der Mobilitätsverwaltung, die bezüglich einer mobilen Instanz im Netz gehalten werden, können in zwei Kategorien [FSY90] eingeteilt werden: Ortsinformationen (*location information*) und Teilnehmerprofil (*personal profile information*).³ Ortsinformationen helfen, einen Teilnehmer zu lokalisieren und eine Kommunikationsbeziehung zu ihm aufzubauen. Das Teilnehmerprofil enthält dienst- und vertragsspezifische Daten, so zum Beispiel Art und Leistungsumfang abonniertes Dienste sowie Parameter von Zusatzdiensten und zur Authentifizierung. Datensätze mobiler Instanzen werden typischerweise mit Hilfe eindeutiger Bezeichner unter-

2. Gegebenenfalls müssen Mechanismen zur Korrektur der Folgen der Verwendung von veralteter Information vorhanden sein.
3. Im RACE-Projekt MONET [MON95a] wurde für UMTS ein deutlich komplexeres Datenmodell erarbeitet. Es wird zwischen Benutzerdaten, Endgerätedaten, Dienstanbieterdaten und Teilnehmerdaten unterschieden. Die Benutzerdaten werden in vier Unterklassen gegliedert: *subscription*, *routing*, *registration* und *session*. Für die vorliegende Arbeit ist es jedoch ausreichend, lediglich zwei Kategorien zu unterscheiden, welche durch zwei wesentliche, unterschiedliche Zugriffsmuster charakterisiert sind.

schieden. In GSM sind dies die IMSI und die MSISDN, welche dem Teilnehmer permanent zugeordnet sind, sowie die TMSI und die MSRN, die temporär vergeben werden.⁴ Entsprechend ihrer unterschiedlichen Verwendung bieten sich für die beiden Datenkategorien unterschiedliche Datenhaltungsstrategien an.

Das Teilnehmerprofil wird zumeist am momentanen Aufenthaltsort des Teilnehmers benötigt. In GSM wird das Teilnehmerprofil einerseits zentral und dauerhaft im HLR gespeichert, andererseits wird eine temporäre Kopie im VLR des momentanen Aufenthaltsorts teilnehmernah gehalten und bei einem VLR-Wechsel vom HLR erfragt. Denkbar wäre auch, das Teilnehmerprofil oder eine Kopie davon im Endgerät bzw. im Teilnehmer-Identifikationsmodul abzulegen. Da die Kosten für die Kommunikation zwischen Endgerät und Netz insbesondere bei schnurloser Anbindung jedoch meist relativ hoch sind und da im Netz Daten aus dem Teilnehmerprofil auch bei Nichterreichbarkeit des Endgeräts benötigt werden, ist es im allgemeinen effizienter und effektiver, das Teilnehmerprofil im Netz zu halten und dort gegebenenfalls von Ort zu Ort weiterzureichen. Unterteilt man die Daten im Teilnehmerprofil etwas feiner, etwa in Parameter von Zusatzdiensten für kommende Rufe und Parameter von Zusatzdiensten für gehende Rufe, so können auch hier jeweils unterschiedliche Datenhaltungsstrategien verfolgt werden.

Auf die Ortsinformationen dagegen wird von verschiedenen Stellen aus zugegriffen, nämlich von den Stellen aus, an denen eine Adreßumsetzung von einer permanenten Adresse einer Instanz auf eine temporäre Adresse stattfindet. Der Platzierung von Daten und Datenkopien sowie den Verfahren zum Ausfindigmachen und Aktualisierung der Daten kommt beim Entwurf einer effizienten Mobilitätsverwaltung damit eine große Bedeutung zu. Im Folgenden soll die Datenhaltung im allgemeinen – also sowohl für das Teilnehmerprofil als auch für Ortsinformationen – diskutiert werden.

In bestehenden und vorgeschlagenen Systemen (siehe Kapitel 3.5) basiert die Datenhaltung zumeist auf mehreren Datenbanken an verschiedenen Orten, so daß eine verteilte Datenhaltung vorliegt. Grundsätzlich lassen sich dabei Systeme unterscheiden, bei denen die Verteilung der Daten den Protokollen der Mobilitätsverwaltung bekannt ist, und Systeme, bei denen ein verteiltes Datenbanksystem die Verteilung der Daten vor den Protokollen der Mobilitätsverwaltung verbirgt. Im ersten Fall sind die Protokolle in der Lage, unter Verwendung des Wissens über den Zweck von Datenoperationen, die Verteilung der Daten zu steuern. Im zweiten Fall werden die Mobilitätsverwaltungsprotokolle von dieser Aufgabe befreit, was eine Vereinfachung derselben zur Folge hat und was flexiblere Datenhaltungsstrategien im verteilten Datenbanksystem ermöglicht, die dann von den Mobilitätsverwaltungsprotokollen nicht berücksichtigt werden müssen. Der Einsatz klassischer verteilter Datenbanksysteme [ÖV91, BG92] bietet sich hier im allgemeinen nicht an. Diese sind für komplexe Anfragen und eine Vielzahl von Benutzern, die von verschiedenen Orten sowohl lesend als auch schreibend auf

4. Daneben gibt es in GSM noch andere eindeutige Bezeichner, wie zum Beispiel die Geräteerkennung (permanent) oder die lokale Teilnehmererkennung (temporär), die hier jedoch der Einfachheit halber nicht berücksichtigt werden.

Daten zugreifen, ausgelegt und müssen dabei jederzeit die Konsistenz der Daten sicherstellen ohne jedoch strengen Anforderungen an die Antwortzeiten zu unterliegen. Der Einsatz spezieller Datenbanksysteme zur Mobilitätsverwaltung wird dagegen durch eine Reihe von spezifischen Randbedingungen motiviert:

- ❑ Anfragen an die Datenbank sind sehr einfacher Natur. Mit Hilfe eines eindeutigen Suchschlüssels wird auf genau einen Datensatz (lesend oder schreibend) zugegriffen.
- ❑ Datenänderungen werden im allgemeinen nur von einer Instanz aus durchgeführt.
- ❑ Bestimmte Daten (z.B. im Teilnehmerprofil) werden nur lokal am momentanen Aufenthaltsort des Teilnehmers nachgefragt. Der momentane Aufenthaltsort verändert sich jedoch im Lauf der Zeit.
- ❑ Antwortzeiten sind enge Grenzen gesetzt, da sie die Dienstgüte beeinflussen.
- ❑ Datenbankzugriffe sind in einige wenige Protokollszenarien eingebettet, die vom Teilnehmerverhalten beeinflußt sind. Informationen über den Zweck dieser Abläufe machen Voraussagen über Art und Ort zukünftiger Anfragen möglich.
- ❑ Bestimmte Daten, wie zum Beispiel die temporär vergebene Aufenthaltsrufnummer, sind nicht statisch vorhanden, sondern müssen bei Bedarf zugewiesen und eindeutig gehalten werden.
- ❑ Die Konsistenz verschiedener Kopien eines Datensatzes hat eine geringere Wichtigkeit, da veraltete Information in der Regel durch Scheitern von Protokollabläufen erkannt wird und eine erneute Anfrage an das Datenbanksystem gerichtet werden kann.⁵

Ein Teil dieser Randbedingungen ermöglicht drastische Vereinfachungen der Funktionalität des verteilten Datenbanksystems. Das Umsetzen komplexer Anfragen in eine effiziente Folge von verteilten Verarbeitungsschritten (*query optimization*) entfällt. Mechanismen zur Behandlung von Nebenläufigkeit bei Änderungen werden nicht unbedingt benötigt. Einfache Transaktionsprotokolle können verwendet werden, da Konsistenz verschiedener Kopien eines Datensatzes nicht unter allen Umständen erzwungen werden muß.

Es werden durch die Randbedingungen jedoch auch besondere Anforderungen an das verteilte Datenbanksystem gestellt. Zum einen ändern sich die Aktivitätszentren, was die schnelle Verlagerung von Datensätzen notwendig macht. Zum anderen kommen Anfragen von beliebigen Stellen, die alle bestimmten Anforderungen an die Antwortzeiten unterliegen. Dem schnellen Ausfindigmachen von Datensätzen kommt dabei eine besondere Bedeutung zu. Schließlich ist eine Interaktion mit Mechanismen, die auf temporärer Basis Kennungen vergeben, notwendig.

Einige der Randbedingungen legen eine Durchbrechung der in [Dat98] aufgestellten Grundregeln für verteilte Datenbanksysteme nahe: „*To the user, a distributed system should look exactly like a **nondistributed** system.*“ Die Tatsache, daß die Mobilitätsverwaltungsprotokolle – die „Benutzer“ der verteilten Datenbanken – implizites Wissen über die verteilte Natur des

5. Die Auswirkungen von veralteter Information auf die Antwortzeiten der Mobilitätsverwaltung, und damit auf die Dienstgüte, müssen für jedes Protokoll in Abhängigkeit vom Benutzerverhalten einzeln bewertet werden.

Systems und den Zweck von Protokollenszenarien haben, ermöglicht es ihnen, dem verteilten Datenbanksystem Anstöße zur Verlagerung und Replizierung von Datensätzen zu geben. Es sei hier angemerkt, daß dies nicht unbedingt die ebenfalls in [Dat98] aufgestellten *zwölf Zusatzregeln für verteilte Datenbanksysteme* verletzt – Regel 4 fordert Ortsunabhängigkeit (*location independence*), Regel 6 Replizierungsunabhängigkeit (*replication independence*), – da Anfragen nach wie vor unabhängig von Ort und Kopienanzahl eines Datensatzes gestellt werden können. Darüberhinaus gibt der Benutzer jedoch zusätzliche Hinweise, die den Datenverwaltungsalgorithmen im verteilten Datenbanksystem Hilfestellung leisten können.

Im RACE-II-Projekt MONET [MON95a] geht man hier noch einen Schritt weiter. Dort sind die Schnittstellen zwischen der Dienststeuerung und den verteilten Datenbanken anwendungsspezifisch definiert. Die Schnittstelle zwischen Mobilitätsverwaltung und verteiltem Datenbanksystem (*Distributed Database*, DDB) besteht aus spezifischen Anfrageprimitiven wie zum Beispiel `O_Loc_Request()` zum Erfragen einer Aufenthaltsrufnummer und `Loc_Update_Request()` zur Aufenthaltsortsaktualisierung. Erst innerhalb der DDB setzt ein sogenannter *Query Handler* diese Primitive in generischere Datenbankoperationen um. Diese sind nach [MON94, Law95]: Auslesen, Erstellen, Aktualisieren, Löschen, Bewegen und Verriegeln. Dabei ist jedoch anzumerken, daß die in [MON94] vorgeschlagenen Datenhaltungsstrategien für einen Datensatz nur maximal eine weitere Kopie vorsehen.⁶

Verzeichnissysteme wie das *Domain Name System* im Internet und das OSI-Verzeichnissystem gemäß den ITU-T Empfehlungen der X.500-Serie⁷ [X.500] erfüllen einen Teil der Anforderungen, die durch die dargestellten Randbedingungen gegeben sind. Solche Verzeichnissysteme sind bisher jedoch zumeist auf häufige Anfragen und seltene Änderungen optimiert und werden deshalb für Mobilitätsverwaltung kaum in Betracht gezogen. Eine Ausnahme stellt [BHJK96] dar, wo X.500 mit angepaßten Replizierungsmechanismen zur Mobilitätsunterstützung von DECT-Endgeräten vorgeschlagen wird, ohne jedoch die Leistungsfähigkeit des Gesamtsystems zu bewerten. Insgesamt stellen die aktuellen Empfehlungen der X.500-Serie sicherlich eine interessante Basis für zukünftige Verzeichnissysteme zur Unterstützung der Mobilitätsverwaltung dar. Aus diesem Grund werden an mehreren Stellen in den folgenden Unterkapiteln Teilaspekte von X.500 kurz beleuchtet.

3.3.1 Plazierung von Daten

Betrachtet man alle Daten, die einer mobilen Instanz zugeordnet werden, als einen Datensatz und die Menge aller dieser Datensätze als eine Datenbanktabelle, so kann man die Verteilung

-
6. Die Datenhaltung innerhalb eines UMTS-Teilgebiets ist in diesem Punkt mit GSM vergleichbar, da Daten in einem sogenannten *resident node* bzw. in einem sogenannten *visited node* gehalten werden, die dem HLR und dem VLR ähnlich sind. Die Adressierung bzw. das Ausfindigmachen dieser Knoten unterscheidet sich jedoch drastisch von GSM, da dazu auf eine hierarchische Verzeichnisstruktur zurückgegriffen wird.
 7. Im folgenden steht der Begriff „X.500“ stellvertretend für das OSI-Verzeichnissystem gemäß den ITU-T Empfehlungen der X.500-Serie.

der Daten sowohl als horizontale als auch als vertikale Fragmentierung der Datenbanktabelle bezeichnen. Vertikale Fragmentierung beschreibt die Aufspaltung einer Tabelle in unterschiedliche Datenkategorien, wie zum Beispiel in Ortsinformationen und Teilnehmerprofil, horizontale Fragmentierung beschreibt die Trennung verschiedener Datensätze einer Tabelle [ÖV91].⁸ Für die Platzierung eines solchen Fragments gibt es nun verschiedene Möglichkeiten:

- ❑ Platzierung an einer oder an mehreren Stellen
- ❑ feste oder adaptive Platzierung
- ❑ zentrale oder dezentrale Steuerung der Platzierung.

Werden bestimmte Daten, wie zum Beispiel Ortsdaten, an verschiedenen Stellen benötigt, so kann es sinnvoll sein, diese auch an mehreren Stellen zu halten. Für lesende Zugriffe wird damit der Kommunikationsaufwand und die Antwortzeit verringert. Für Aktualisierung und Konsistenzhaltung entsteht dagegen ein höherer Kommunikationsaufwand.

Sind Zugriffsraten auf die Fragmente zeitlich konstant, so bietet sich eine feste Platzierung der Fragmente an. Ist bekannt, wie häufig und in welcher Weise von bestimmten Stellen auf Fragmente zugegriffen wird, so liegt das klassische, NP-vollständige *File Allocation Problem* vor [LL83], bei dem Ort und Anzahl von Kopien eines Fragments optimiert werden, so daß die Kosten für Lesen, Aktualisieren und Speicherplatz minimiert werden. Treffen die obigen Annahmen nicht zu, so ist auch eine adaptive Platzierung von Kopien denkbar. Dabei sind sowohl die Kosten für die Duplizierung und Verlagerung von Fragmenten als auch der u. U. erhöhte Aufwand zum Ausfindigmachen der Kopien zu berücksichtigen.

Bei adaptiver Platzierung unterscheidet [Ape83] zwischen zentraler und dezentraler Steuerung. Bei zentraler Steuerung wird vom globalen Datenbankmanagementsystem über die Platzierung aller Kopien entschieden, bei dezentraler Steuerung dagegen werden Kopien von einzelnen Datenbankknoten bei Bedarf angefordert und gegebenenfalls auch wieder gelöscht. Im letzteren Fall wird eine Kopie häufig zum exklusiven Gebrauch durch einen Datenknoten erstellt. Solch eine Kopie wird dann im allgemeinen als *private Kopie* bezeichnet. Die dezentrale Steuerung reagiert im allgemeinen schneller auf Änderungen der Zugriffsmuster, während die zentrale Steuerung bei stabileren Zugriffsmustern besser die globalen Kosten für Zugriffe, Aktualisierung und Verlagerung optimieren kann.

Seit der jüngsten Version der Empfehlungen [X.500] sieht X.500 die Replizierung von Daten vor. In der ITU-T Empfehlung X.525 [X.525] wird dazu das sogenannte *shadowing*-Verfahren spezifiziert. Wird eine Kopie direkt aus den Primärdaten (*master copy*) erstellt, so spricht man von *primary shadowing*. Wird dagegen eine Kopie aus einer Kopie erzeugt, so spricht man von *secondary shadowing*. Zwischen dem Lieferanten der Daten und dem Empfänger der Kopie wird dabei das Aktualisierungsverfahren für die Kopie festgelegt. Das Primärdaten und die

8. Betrachtet man die Platzierung verschiedener Datenfragmente als unabhängig, so erübrigen sich Differenzierungen wie in [LL97b, LL97a], wo zum Beispiel zwischen *full* und *partial replication scheme* unterschieden wird.

erstellten Kopien sind jedoch nicht gleichwertig. Während auf die Primärdaten lesend und schreibend zugegriffen werden kann, dienen die Kopien nur zur Beschleunigung von Lesezugriffen. Aktualisierungen werden direkt auf den Primärdaten durchgeführt, wozu jede Kopie eine Referenz auf ein Primärdatum besitzt. [Kes96] unterscheidet hierarchische Replizierung und gleichrangige Kopien. *Shadowing* ist ein Beispiel für hierarchische Replizierung, während bei gleichrangigen Kopien Lese- und Schreibzugriffe gleichberechtigt auf alle Kopien möglich ist.

Einen besonderen Fall stellen in diesem Zusammenhang sogenannte *Cache*-Kopien dar. Dies sind Kopien, die lokal erstellt werden, ohne daß ihre Existenz darüberhinaus bekannt gemacht wird. Daraus folgt, daß sie über Änderungen nicht explizit informiert werden und daß über ihre Aktualisierung oder Löschung lokal entschieden wird. Bei X.500 wird zwar zwischen *shadowing* und *caching* unterschieden, die Empfehlung X.525 beschäftigt sich jedoch ausschließlich mit *shadowing*-Verfahren.

Ein weiterer wichtiger Aspekt der Datenplatzierung ist die Begrenzung des Verfügbarkeitsbereiches der Daten. Bestimmte Daten dürfen u. U. nur für bestimmte Personen oder Instanzen zugänglich sein. Dazu ist es sinnvoll, den Verfügbarkeitsbereich von Daten zu beschränken. Der Verfügbarkeitsbereich geheimer Schlüssel kann beispielsweise auf das Teilnehmer-Identifikationsmodul und ein Berechtigungszentrum beschränkt sein. Ebenso kann die Verfügbarkeit auf Zugriffe aus dem Heimatnetz beschränkt sein und die Daten damit für fremde Netze und Netzbetreiber unzugänglich gemacht werden. In [Gun97] wird eine Lösung zur Realisierung von Benutzermobilität vorgeschlagen, bei der Dienstlogik und Dienstdaten sich ausschließlich im Verfügbarkeitsbereich des Teilnehmers befindet und das Telekommunikationsnetz lediglich als „transparentes Verbindungsnetz“ fungiert. Dabei wird mit Hilfe des ISDN-Zusatzdienstes Rufumlenkung (*call deflection*) ein Ruf auf Anweisung des gerufenen Endgeräts von der Heimatvermittlungsstelle eines Teilnehmers zu einem anderen Endgerät weitergeleitet. Ähnliche Konzepte mit einer Dienststeuerung durch im Verfügbarkeitsbereich des Teilnehmers befindliche Steuerknoten ließe sich auch im Rahmen des „Intelligenten Netzes“ realisieren, so daß beispielsweise ein *Personal Computer* als SCP agieren könnte, der über eine entsprechende Schnittstelle Abläufe im Netz steuern oder beeinflussen könnte.

3.3.2 Ausfindigmachen von Daten

Soll auf ein Datenfragment oder auf eine Kopie eines Datenfragments zugegriffen werden, so muß der Ort, an dem sich dieses Datenfragment befindet, ausfindig gemacht (lokalisiert) werden. Hierzu gibt es drei mögliche Varianten, die meist auch kombiniert werden:

- ❑ Der Ort eines Datenfragments ist implizit bekannt.
- ❑ Der Ort eines Datenfragments wird durch Befragen eines Datenverzeichnisses ermittelt.
- ❑ Die Datenbankanfrage wird von Datenbankknoten zu Datenbankknoten weitergereicht, bis sie beantwortet werden kann. Dabei werden häufig Verweiseiger verwendet.

Oftmals ist der Ort eines Datenfragments implizit bekannt. Weiß man beispielsweise, daß bestimmte Daten grundsätzlich am momentanen Aufenthaltsort eines Teilnehmers gespeichert werden, so kann dort lokal auf diese Daten zugegriffen werden. In GSM ist das HLR eines Teilnehmers (und damit der Ort der dort gespeicherten Daten) sowohl aus der IMSI als auch aus der MSISDN – also aus den Suchschlüsseln – ableitbar. Das HLR wird im Signalisiersystem Nummer 7 direkt mit Hilfe der Suchschlüssel adressiert. Eine Änderung der Zuordnung von Suchschlüssel zu HLR kann dabei nur durch Änderung der *Global Title Translation* im SCCP erfolgen.

Klassische verteilte Datenbanken besitzen einen sogenannten Systemkatalog bzw. ein Datenverzeichnis, in dem Metadaten⁹ gehalten werden, zu denen auch der Ort eines Datenfragments gehören kann. [BG92] unterscheidet drei Realisierungsmöglichkeiten:

- zentral,
- voll vervielfacht (*fully replicated*) und
- verteilt.

Bei einer zentralen Lösung muß der Ort eines Datenfragments von einer zentralen Instanz erfragt werden. Bei voller Vervielfachung hat jeder Datenbankknoten die Information über den Ort aller Datenfragmente im System. Bei einer verteilten Lösung ist das Datenverzeichnis selbst wieder ein verteiltes Datenbanksystem, in dem Replizierung und Lokalisation von Metadaten durch die hier betrachteten Mechanismen unterstützt wird.

Wird eine Anfrage an einen Datenbankknoten gestellt und kann dieser sie nicht direkt beantworten, so bietet sich das Weiterreichen dieser Anfrage an. Diese Variante ergänzt häufig die beiden anderen in der Form, daß der erste Knoten, an den eine Anfrage gerichtet wird, implizit oder mit Hilfe des Datenverzeichnisses ermittelt wird. Für das Weiterreichen der Anfrage lassen sich verschiedene Techniken unterscheiden:

- Verweisen (*Referral*),
- Verketteten (*Chaining*) und
- Abgeben (*Passing*).

Beim *Verweisen* wird eine erfolglose Anfrage mit einem Verweis auf einen anderen Datenbankknoten beantwortet, der als nächster befragt werden soll. Beim *Verketteten* stellt der befragte Datenbankknoten selbst eine Anfrage an einem anderen Knoten. Das Ergebnis dieser Anfrage liefert er an den ursprünglichen Anfrager zurück. Beim *Abgeben* dagegen wird beim Weiterreichen der Anfrage eine Referenz auf den ursprünglichen Anfrager mitgegeben, so daß das Resultat direkt vom bearbeitenden Datenbankknoten an den ursprünglichen Anfrager zurückgegeben werden kann. Abbildung 3-4 veranschaulicht die drei Techniken. Eine Sonderform des Verkettens stellt das *Multicasting* bzw. *Multichaining* dar, bei dem ein Datenbankknoten, der eine Anfrage nicht bearbeiten kann, diese parallel oder sequentiell an mehrere

9. Daten über Daten.

andere Knoten weiterreicht und aus den erhaltenen Resultaten eine Antwort an den Anfrager erstellt. Die Techniken *Referral*, *Chaining* und *Multichaining* sind bereits bei X.500 vorgese- hen [X.518]. *Passing* wurde im RACE-Projekt MONET zusätzlich vorgeschlagen. In [MON95b, Law95] werden die verschiedenen Techniken qualitativ bewertet.

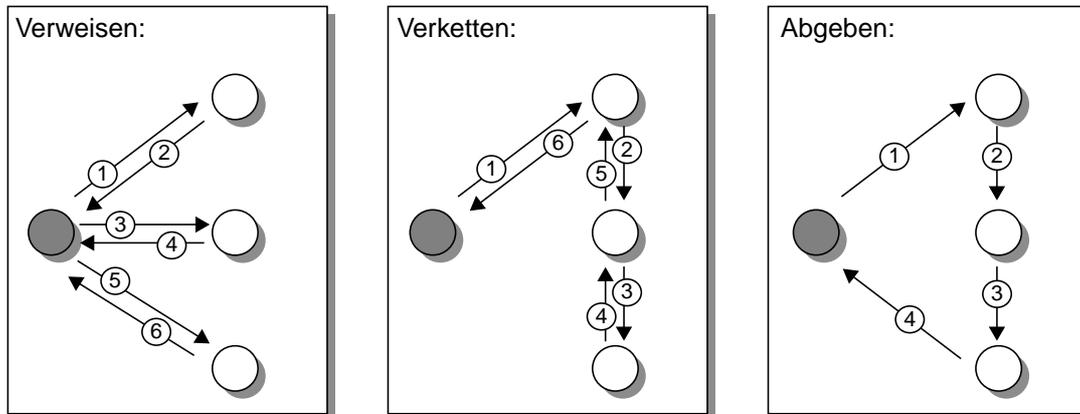


Abbildung 3-4: Drei verschiedene Techniken zum Weiterreichen von Datenbankankfragen (*Verweisen*, *Verketteten* und *Abgeben*), veranschaulicht mit jeweils drei Datenbankknoten.

Für alle Weiterreichungstechniken gilt, daß die Instanz, auf die verwiesen wird, ermittelt werden muß. Dies geschieht entweder explizit durch das Vorhandensein von Verweiszeigern oder implizit auf der Grundlage von Regeln, die Wissen über die Methodik der Datenverteilung und die Struktur des verteilten Datenbanksystems haben können. So kann beispielsweise in einem hierarchisch organisierten verteilten Datenbanksystem mit Baumstruktur eine Anfrage an den nächsten Knoten in Richtung Wurzel weitergereicht werden. Auch die Kombination von Verweiszeigern und Regeln ist denkbar. So kann beispielsweise ein Zeiger auf eine Gruppe von Datenbankknoten im Datenbanksystem verweisen, die mit Hilfe der *Multichaining*-Technik befragt werden. Bei parallelem *Multichaining* kann man dann – analog zum Vorgehen beim Ausrufen von Teilnehmern über die Luftschnittstelle von GSM – von *Paging* im verteilten Datenbanksystem sprechen. Bei sequentiell *Multichaining* kann man von *Polling* sprechen. Ein hybrides Verfahren, bei dem sequentiell *Paging* in Teilen einer Gruppen erfolgt ist ebenso denkbar. Dies wurde schon mehrfach für die Luftschnittstelle [Pla94, HA95, Tab95, AHL96] und auch für das Festnetz [TN95] vorgeschlagen.

Verfahren, welche das Ausfindigmachen von Daten mit Hilfe von Regeln und/oder Verweiszeigern durchführen, haben in der Regel einen höheren Bearbeitungsaufwand zur Folge. Durch ihre Fähigkeit, Ortsinformation hierarchisch zu verteilen, ermöglichen sie jedoch häufig einen verringerten Kommunikationsaufwand. Dies gilt sowohl für das Ausfindigmachen von Daten als auch für die Änderung ihrer Plazierung.

3.3.3 Datenaktualisierung

Unter dem Begriff Datenaktualisierung sollen hier zwei unterschiedliche Vorgänge zusammengefaßt werden: zum einen die eigentliche Änderung eines Datums und zum anderen die Änderung der Plazierung eines Datenfragments. Diese Zusammenfassung motiviert sich aus der

Tatsache, daß bei der Ortsänderung einer mobilen Instanz in einem Kommunikationsnetz im allgemeinen Vorgänge beider Art stattfinden. Darüberhinaus ist die Änderung der Platzierung eines Datenfragments zumeist mit der Änderung von Metadaten verbunden, so daß auch dieser Vorgang eine Datenaktualisierung darstellt.

Ist ein Datenfragment nur an einer Stelle vorhanden (bzw. gibt es nur ein zentrales Datenverzeichnis), so ist eine Datenaktualisierung relativ einfach: das zu ändernde Datum muß ausfindig gemacht und die Änderung vorgenommen werden. Anders liegt der Fall, wenn von einem Datenfragment mehrere Kopien existieren (bzw. das Datenverzeichnis vervielfacht oder verteilt ist).

Bei hierarchischer Replizierung wird das Primärdatum ausfindig gemacht und an ihm die Änderung vorgenommen. Die Aktualisierung der nachgeordneten Kopien sowie der Kopien von Kopien kann dann auf unterschiedliche Weise erfolgen. Bei X.500 wird dazu zwischen dem Lieferanten der Daten und dem Halter einer Kopie ein Aktualisierungsverfahren festgelegt (*shadowing agreement*).

Bei gleichrangigen Kopien wird eine der Kopien ausfindig gemacht und an ihr die Änderung vorgenommen. Es ist dann die Aufgabe des Datenbankknotens an dem sich die Kopie befindet, die anderen Kopien zu aktualisieren. Im einfachsten Fall besitzt jede Kopie eine Liste aller anderen Kopien und aktualisiert diese bei Änderung. Effizienter kann es sein, wenn eine Aktualisierung nur an einen Teil der anderen Kopien weitergegeben wird und diese dann die Aktualisierung nach deterministischen Regeln an weitere Kopien weiterreichen bis alle Kopien aktualisiert sind. Es ist hier auch denkbar, zwischen Paaren von Kopien individuell ein Aktualisierungsverfahren festzulegen. Alternativ werden in [Kes96] zufallsgesteuerte Verfahren zur Weitergabe von Aktualisierungen untersucht, darunter das „*Anti-Entropy-Verfahren*“, bei dem jede Kopie sich immer wieder mit zufällig ausgewählten anderer Kopien abgleicht, so daß die Verbreitung der Aktualisierungen ähnlich wie die Verbreitung von Epidemien verläuft, sowie das „*Rumor-Spreading-Verfahren*“ bei dem jede Kopie, die eine Aktualisierung erfährt, sich mit einer Folge anderer Kopien abgleicht und nach jedem Abgleich – u. U. abhängig davon, ob der Empfänger bereits aktualisiert war – zufallsgesteuert das Abgleichen beendet. In [Kes96] wird gezeigt, daß zufallsgesteuerte Verfahren „nur dann in Betracht kommen, wenn eine große Anzahl“ von Replikaten „abzusehen ist“ und „Robustheit gegen Fehlersituationen“ wichtiger ist als „Aktualität und Konsistenz der Daten“. Das „*Rumor-Spreading-Verfahren*“ mit weniger als 20 Replikaten wird in [Kes96] als extrem unzuverlässig bezeichnet.

Generell lassen sich die Aktualisierungsverfahren danach unterscheiden, ob eine Änderung von der geänderten Kopie (bzw. dem geänderten Primärdatum) an andere Kopien weitergeleitet wird oder ob eine Änderung erst auf Anfrage übermittelt wird. X.500 unterscheidet hier *supplier update mode* und *consumer update mode*. [Kes96] verwendet die Begriffe *push* und *pull*. Bei X.500 werden im *supplier update mode* Kopien bei Änderung des Primärdatums sofort aktualisiert, während im *consumer update mode* periodischen Anfragen und/oder durch besondere Ereignisse (wie z. B. das Fehlschlagen eines Rufaufbaus) ausgelöste Anfragen vor-

gesehen sind. Datenabgleich nach deterministischen Regeln kann prinzipiell periodisch oder durch Änderungen bzw. besondere Ereignisse ausgelöst erfolgen. Im letzteren Fall kann man noch sofortigen Abgleich oder Abgleich zu festgelegten Zeitpunkten unterscheiden.

Ein wesentlicher Aspekt der Datenaktualisierung ist die Sicherung der Datenkonsistenz. In klassischen Datenbanksystemen sorgen Transaktionsprotokolle dafür, daß beim Ausführen einer ändernden Datenbankoperation immer ein Übergang zwischen dauerhaften, konsistenten Zuständen erfolgt. Mechanismen zur Steuerung nebenläufiger Operationen stellen dabei sicher, daß die Nebenläufigkeit von Operationen nicht zu gegenseitigen Beeinträchtigungen führt. In verteilten Datenbanksystemen mit einer Vielzahl an Kopien und nebenläufigen Operationen bedeutet die Sicherung der Datenkonsistenz und die Vermeidung von Verklemmungen einen erheblichen Aufwand. Wie auf Seite 60 dargelegt, unterliegt die Datenhaltung zur Mobilitätsverwaltung besonderen Randbedingungen. Man kann davon ausgehen, daß zwischen verschiedenen Datensätzen im allgemeinen keine Verknüpfungen bestehen und damit Änderungen daran unabhängig vorgenommen werden können, daß Datenänderungen zumeist nur durch eine einzige Instanz angestoßen werden und daß nicht alle Kopien eines Datensatzes zu jeder Zeit dieselben Werte haben müssen. Daher lassen sich sehr viel einfachere Protokolle zur Aktualisierung entwerfen, die schnellere Antwortzeiten und geringeren Kommunikationsaufwand zur Folge haben. In [Leu97] wird beispielsweise für vergleichbare Anwendungen das sogenannte *Primary-Writer-Protocol* vorgeschlagen, bei dem Änderungen zuerst an den Primärdaten vorgenommen werden und dann die konsistente, zeitlich versetzte Änderung von Kopien mit Hilfe von Folgenummern sichergestellt wird. Es werden dabei veraltete Daten und deren Folgen in Kauf genommen,¹⁰ andererseits jedoch Blockierungen, Verklemmungen und lange Antwortzeiten vermieden. Allgemein gilt, daß man sich in einem System, in dem nur eine einzige Instanz Änderungen anstößt, mit Hilfe von Folgenummern für Änderungsoperationen gegen unerwünschte Auswirkungen von Verlusten oder Reihenfolgeänderungen derselben schützen kann. Bei der Ortsaktualisierung können damit beispielsweise aktuellere Daten akzeptiert und Verspätungen oder vorherige Verluste ignoriert werden. Bei komplexeren Operationen kann damit die Reihenfolge ihrer Ausführung sichergestellt werden. Bei mehreren ändernden Instanzen werden zu ähnlichen Zwecken häufig auch Zeitstempel und synchronisierte Uhren verwendet [Kes96].

Während bei der Verwendung eines Datenverzeichnisses die Änderung der Platzierung eines Datenfragments eine Aktualisierung des Datenverzeichnisses nach sich zieht, müssen bei der Verwendung von Weiterreichungsverfahren Verweiszeiger verändert werden. Je nach Verfahren kann es dann an verschiedenen Stellen zum Erzeugen, Verändern oder Löschen derselben kommen. Nimmt man beispielsweise an, daß der Ort eines Datenfragments durch eine Kette von Verweiszeigern bezeichnet wird, so kann eine Änderung der Platzierung des Datenfrag-

10. Wie bereits vermerkt, können veraltete Informationen in der Regel durch Scheitern von Protokollabläufen erkannt werden und daraufhin entsprechende Maßnahmen ergriffen werden.

ments die Kette einfach um einen weiteren Verweiszeiger vom alten zum neuen Ort verlängert werden, oder aber die gesamte Kette kann neu konfiguriert werden, so daß durch Löschen, Verändern und Hinzufügen eine – u. U. kürzere – aktualisierte Verweiszeigerkette entsteht.

3.4 Verbindungsweitergabe

Bewegt sich eine mobile Instanz, während sie Kommunikationsbeziehungen unterhält, so muß sich gegebenenfalls die Nutzverkehrsführung ändern (*handover*). Dabei ist man bestrebt, solch eine Verbindungsweitergabe möglichst nahtlos (*seamless*) zu gestalten, d. h. so, daß der Benutzer diese nicht wahrnimmt. Man unterscheidet zwei Verbindungsweitergabe-Techniken: *hard handover* und *soft handover*. Bei einem *soft handover* wird Nutzverkehr zeitweise vervielfacht, auf verschiedenen Wegen geführt und an geeigneter Stelle wieder kombiniert. Beim *hard handover* dagegen wird der Nutzverkehr lediglich von einem Weg auf einen anderen umgelenkt. Ein *soft handover* erfolgt im allgemeinen nahtlos. Bei einem *hard handover* bedarf es (in Abhängigkeit vom transportierten Dienst) spezieller Anstrengungen, um ihn nahtlos durchführen zu können.

Ein großer Teil der Verbindungsweitergaben wird im allgemeinen im Zugangsnetz durchgeführt. Erst eine Veränderung des Netzzugangspunktes bedarf der Unterstützung von Mobilitätsverwaltungsmechanismen im Kernnetz. Ändert eine mobile Instanz ihren Netzzugangspunkt, so ist es die Aufgabe der Mobilitätsverwaltung im Kernnetz, bestehende Verbindungen aufrecht zu erhalten. Dazu bietet sich an, entweder Verbindungen vom zuletzt besuchten Ort zum neuen Netzzugangspunkt lokal zu verlängern, oder aber die Verbindungsführung weiterreichend zu rekonfigurieren. Bei einer weiterreichenden Rekonfigurierung ist jedoch darauf zu achten, daß Steuerbeziehungen, die Instanzen zur Dienststeuerung zu den Verbindungen unterhalten, nicht abreißen. Heutige Systeme unterstützen dies jedoch im allgemeinen nicht. Im „Intelligenten Netz“ zum Beispiel müßte dazu die Rufsteuerung von einem SSP an einen anderen SSP verlagert werden können, was die Übergabe von Zustandsinformation und den Transfer von SSF-SCF-Steuerbeziehung einschließen würde. Gegebenenfalls müßte sogar die Kontrolle über einen Ruf von einem SCP an einen anderen SCP übergeben werden. In Annex I von [ETSI96] werden verschiedene Varianten zur Realisierung der Weitergabe von Dienststeuerbeziehungen diskutiert.

In GSM wird eine lokale Verlängerung von Verbindungen durchgeführt. Die beim Verbindungsaufbau zuständige MSC behält bis zum Abbau einer Verbindung die Kontrolle über den Ruf und wird als Anker-MSC bezeichnet. Jede durch eine Verbindungsweitergabe neu hinzukommende MSC verlängert die Verbindungsführung. Lediglich bei Schleifen im verlängerten Teil wird die Verbindungsführung verkürzt. Bei UMTS ist man bestrebt, Verbindungsweitergaben im Zugangsnetz und Verbindungsweitergaben im Kernnetz zu entkoppeln. Dazu soll das Zugangsnetz in die Lage versetzt werden, den Übergangspunkt von Verbindungen ins Kernnetz – und damit den Netzzugangspunkt – ungeachtet des momentanen Aufenthaltsortes einer

mobilen Instanz aufrecht erhalten zu können und unabhängig von Verbindungsweitergaben im Zugangsnetz, eine Verlagerung des Übergangspunktes zum Kernnetz anstoßen zu können. Bei *Mobile IP* erfolgt eine Verbindungsweitergabe implizit, da Pakete zur neuen *care-of address* eines mobilen Rechners geschickt werden, sobald diese beim *Home Agent* bekannt ist. Dabei treten – wie in Kapitel 2.5.1 dargelegt – im allgemeinen massive Paketverluste auf, die bisher nur von Protokollen oberhalb von IP (z.B. von TCP) korrigiert werden können. Im Rahmen der Überlegungen zu *Route Optimization* werden Mechanismen zum Puffern und Weiterleiten von Paketen (*smooth handoff*) angedacht, die solche Paketverluste und deren Auswirkungen auf die Dienstgüte des Nutzverkehrs vermeiden sollen.

3.5 Beispiele

In den vorausgehenden Darstellungen wurden verschiedene Aspekte der Mobilitätsverwaltung abstrakt beschrieben. Nun werden einige konkrete Mobilitätsverwaltungsprotokolle bestehender oder vorgeschlagener Systeme hinsichtlich der beschriebenen Aspekte dargestellt und klassifiziert. Die Art des Zusammenspiels von Wegführung und Zielsuche sowie die Vorgehensweisen bei der Plazierung, beim Ausfindigmachen und beim Aktualisieren von Daten dienen dabei als Kriterien zur Klassifizierung. Darüberhinaus werden zum Teil auch Fragen der Skalierbarkeit und der Datensparsamkeit¹¹ mitbetrachtet.

3.5.1 Datenhaltung in GSM

Die Teilnehmerlokalisierung in GSM wurde bereits in Kapitel 2.3.2 ausführlich beschrieben. Hat ein Ruf seinen Ursprung im Heimatnetz des Gerufenen oder in einem GSM-Netz, welches für den gerufenen Teilnehmer das Dienstmerkmal *Support of Optimal Routing* der GSM-Phase 2+ unterstützt, so erfolgt die Zielsuche vor dem Verbindungsaufbau. Ansonsten wird zuerst eine Verbindung ins Heimatnetz des gerufenen Teilnehmers geführt, von wo dann die Zielsuche initiiert wird. In GSM gibt es zwei Arten von Datenbankknoten: Heimatdateien (HLR) und Besucherdateien (VLR). Jeder Teilnehmer ist genau einem HLR statisch zugeordnet. Eine besuchte Mobilfunk-Vermittlungsstelle ist einem VLR zugeordnet. Für einen einzelnen Teilnehmer ergibt sich eine logische Struktur des Datenbanksystems, bei der das HLR des Teilnehmers die Rolle eines zentralen Knotens spielt, der Daten und Verweiszeiger dauerhaft hält, und bei der die VLRs lediglich temporär Daten teilnehmernah speichern. Abbildung 3-5 zeigt die logische Struktur des teilnehmerspezifischen Datenbanksystems von GSM. Alle Datenbankknoten sind mit Hilfe der *Global Title Translation* im Signalisiersystem Nummer 7 von überall direkt adressierbar.

11. Unter Datensparsamkeit versteht man das Bemühen, möglichst wenige personenbezogene Daten in einem Netz zu halten, um deren möglichem Mißbrauch vorzubeugen. Auf diesem Hintergrund wird in der vorliegenden Arbeit auch auf die nähere Betrachtung von Verfahren verzichtet, die zur Optimierung der Mobilitätsverwaltung Bewegungsprofile von Teilnehmern erstellen.

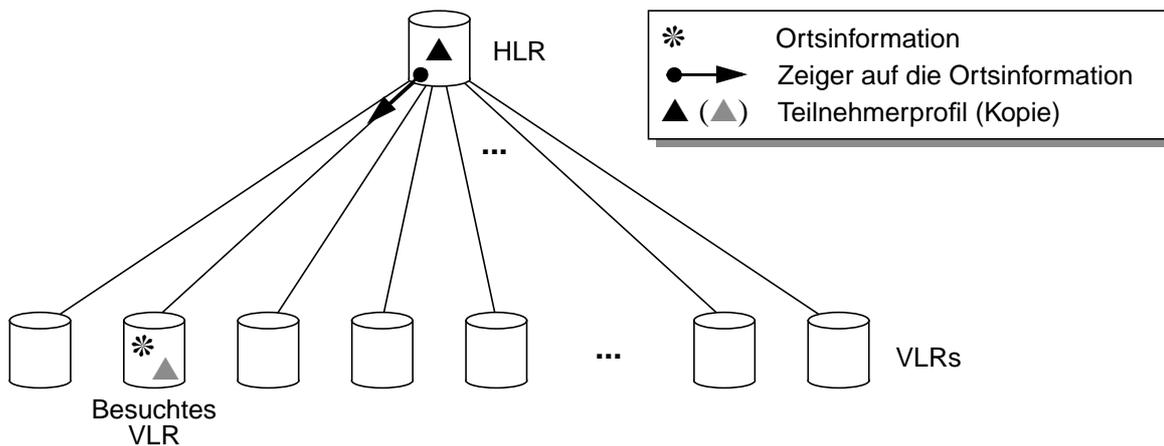


Abbildung 3-5: Logische Struktur des teilnehmerspezifischen Datenbanksystems von GSM, in die exemplarisch die Datenplatzierung für Ortsinformation und Teilnehmerprofil eingezeichnet ist.

Für unterschiedliche Arten von Daten sind unterschiedliche Datenhaltungsstrategien vorgesehen. In GSM wird das Teilnehmerprofil im allgemeinen an zwei Stellen gehalten. Zum einen dauerhaft im HLR und zum anderen adaptiv am jeweiligen besuchten Ort. Bei einem Aufenthaltsortswechsel wird das Teilnehmerprofil vom HLR in das neue VLR kopiert und die Kopie im alten VLR gelöscht. Der Ort des Teilnehmerprofils ist damit für die Mobilfunk-Vermittlungsstelle, an der sich der Teilnehmer befindet und von wo darauf zugegriffen wird, implizit bekannt. Datenänderungen werden am Primärdatum im HLR vorgenommen. Von dort werden dann die Kopien im VLR überschrieben. Authentifizierungsparameter-Sätze werden im Berechtigungszentrum erzeugt und dann von VLR zu VLR weitergereicht, sind also nur an einer Stelle vorhanden. Ihr Ort ist damit wiederum für die Mobilfunk-Vermittlungsstelle, an der sich der Teilnehmer befindet, implizit bekannt.

Die aktuelle Aufenthaltsrufnummer eines Teilnehmers (und damit die Ortsinformation) wird üblicherweise im besuchten VLR erzeugt. Man kann also sagen, daß die Aufenthaltsrufnummer nur an einer Stelle (virtuell) plaziert ist. Die Platzierung ist adaptiv und wird durch das Protokoll gesteuert. Eine Anfrage nach der Aufenthaltsrufnummer wird direkt an das HLR gestellt,¹² welches diese mit Hilfe eines Verweiszeigers an das VLR weiterreicht. Dabei kommt die Weiterreichungstechnik *Verketteten* zum Einsatz. Bei einem Aufenthaltsortswechsel ändert sich die virtuelle Platzierung der Aufenthaltsrufnummer und der Zeiger im HLR (dem „zentralen Datenverzeichnis“) wird aktualisiert. Im HLR ist also jederzeit erkennbar, in welchem VLR ein Teilnehmer momentan registriert ist. Damit könnte eine grobe Aussage über den aktuellen geographischen Aufenthaltsort des Teilnehmers getroffen werden. In Abbildung 3-5 ist die Platzierung von Teilnehmerprofil und Ortsinformation sowie der Verweiszeiger auf die Ortsinformation exemplarisch eingezeichnet.

12. Dieses wird direkt mit Hilfe der öffentlichen Rufnummer des Teilnehmers (als *Global Title*) adressiert.

Die Mobilitätsverwaltung für funkbasierte Mobilkommunikationssysteme wie GSM wurde im Grunde für relativ kleine Netze und eine beschränkte Zahl von Teilnehmern konzipiert, die sich im wesentlichen im Heimat-Mobilfunknetz aufhalten. GSM ermöglicht zwar, daß Teilnehmer auch in fremdem GSM-Netzen erreichbar bleiben („*International Roaming*“), Wegführung und Mobilitätsverwaltung wurden aber für diesen Fall nicht besonders optimiert. Erst mit dem Dienstmerkmal *Support of Optimal Routeing* der GSM-Phase 2+ wird der wachsenden Bedeutung des internationalen „*Roaming*“ Rechnung getragen. Dabei wird jedoch nur die Wegführung verbessert, nicht jedoch die Signalisierung zur Teilnehmerlokalisierung, bei der immer das ferne HLR eines Teilnehmers beteiligt ist.

Für Systeme, die universelle Erreichbarkeit für eine sehr große Zahl von Teilnehmern ermöglichen sollen, wurde Anfang der 90er-Jahre ein Vorschlag erarbeitet, der auf hierarchischer Datenhaltung beruht und dadurch bessere Skalierbarkeit verspricht. Dieser wird im folgenden Abschnitt vorgestellt.

3.5.2 Hierarchische Datenhaltung für universelle Mobilität

[FSY90] befaßt sich mit der Datenhaltung für einen globalen UPT-Dienst, der in einer Vielzahl von kooperierenden Netzen zur Verfügung gestellt werden soll. Dabei werden verschiedene Optionen für die Platzierung und das Ausfindigmachen von Ortsinformation und Teilnehmerprofil diskutiert. Neben Variationen einer GSM-ähnlichen Datenhaltung wird eine hierarchische Datenhaltung vorgeschlagen. Dabei wird für das Teilnehmerprofil davon ausgegangen, daß es wie in GSM sowohl im Heimatnetz, als auch am besuchten Ort gehalten wird, während die Ortsinformation über mehrere Datenbankknoten verteilt vorliegt. Wegführung und Zielsuche wechseln sich dabei ab, so daß die Wegführung durch die Hierarchie der Datenhaltung geprägt wird.

In [Wan93] wird ein sehr ähnlicher Vorschlag präsentiert, der jedoch lediglich die Teilnehmerlokalisierung und nicht die Datenhaltung für das Teilnehmerprofil umfaßt. Darüberhinaus unterscheidet er sich geringfügig bezüglich des Zusammenspiels zwischen Wegführung und Zielsuche. Abbildung 3-6 zeigt die logische Struktur des Datenbanksystems, das in der Form eines Baumes organisiert ist. Die Blätter des Baumes entsprechen den möglichen Aufenthaltsorten eines Teilnehmers und fungieren als Besucherdatei. Die Platzierung des Teilnehmerprofils entspricht der im Vorschlag von [FSY90], die Verteilung partieller Ortsinformation ist beiden Vorschlägen gemein. Für jeden Teilnehmer ist einer der Blätter des Baumes Heimatknoten. Von dort wird entlang der Baumstruktur eine Kette partieller Ortsinformationen zum aktuell besuchten Ort plaziert. Erfragen partieller Ortsinformation ermöglicht Wegführung des Nutzverkehrs zu dem Datenbankknoten, auf den die partielle Ortsinformation verweist.

Die Wegführung wird im Folgenden anhand des Vorschlags in [Wan93] dargestellt: Soll eine Verbindung zum gerufenen Teilnehmer aufgebaut werden, so erfolgt die Wegführung des Nutzverkehrs entlang der Knoten des verteilten Datenbanksystems. An jedem Knoten wird erneut über die weitere Verkehrsführung entschieden. Liegt keine Ortsinformation vor, so fin-

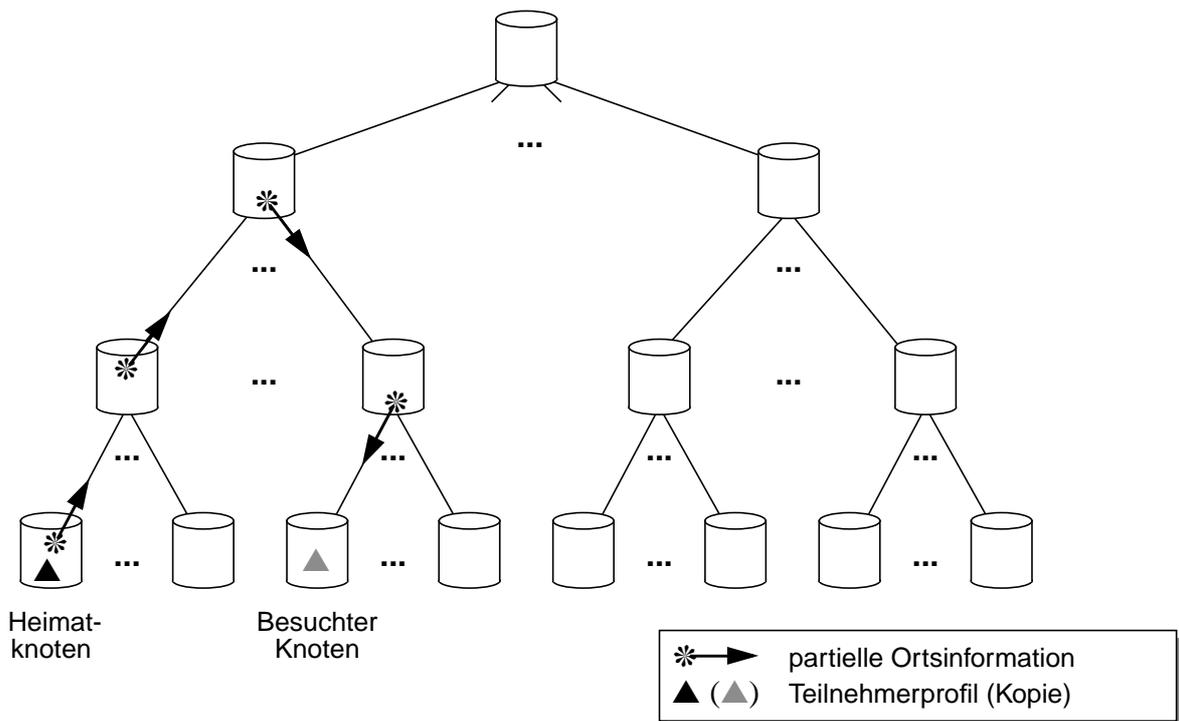


Abbildung 3-6: Struktur des hierarchischen Datenbanksystems nach [FSY90] bzw. [Wan93], in die exemplarisch die Datenplatzierung für partielle Ortsinformation und für das Teilnehmerprofil eingezeichnet ist.

det Wegführung zum nächsten Knoten entlang der Baumstruktur in Richtung des Heimatknotens des Teilnehmers statt. Wird partielle Ortsinformation vorgefunden, so wird sie zur Wegführung zum nächsten Knoten verwendet. Dies geschieht, bis der Knoten erreicht wird, an dem sich der Teilnehmer momentan befindet. Daraus folgt, daß bei diesem System Wegführung und Zielsuche gemeinsam stattfinden. Anders als bei GSM ist deshalb beim Verbindungsaufbau das Mitführen eines Bezeichners für den gesuchten Teilnehmer notwendig. Andererseits erhöht dieses Vorgehen den Aufwand zum Ermitteln des aktuellen geographischen Aufenthaltsorts eines Teilnehmers, da in der Regel weder am Ursprung eines Rufes noch an einem zentralen Punkt vollständige Ortsinformation erlangt werden kann. Es beugt damit Datenmißbrauch in begrenztem Maße vor.

Ortsinformation liegt nicht mehrfach vor, ist jedoch über die verschiedenen Datenbankknoten verteilt. Ihre Platzierung ist adaptiv und wird dezentral gesteuert. Ortsanfragen finden jeweils lokal statt und werden nicht weitergereicht. Im Zusammenspiel mit der Wegführung kann man jedoch das Ausfindigmachen des Aufenthaltsortes als einen Vorgang klassifizieren, bei dem Datenbankankfragen mit Hilfe der Weitergabetechnik *Verweisen* von Datenbankknoten zu Datenbankknoten wandern.

Die Aktualisierung der partiellen Ortsdaten erfolgt vom neuen besuchten Knoten aus. Die Kette der partiellen Ortsinformationen wird umkonfiguriert, indem partielle Ortsinformation gelöscht, verändert oder neu platziert wird. Lediglich Knoten entlang des Pfades vom neuen

besuchten Ort zum alten besuchten Ort sind dabei beteiligt. Abbildung 3-7 illustriert zwei Beispiele für die Aktualisierung der partiellen Ortsinformationen beim Wechsel von einem alten besuchten Knoten zu einem neuen besuchten Knoten.

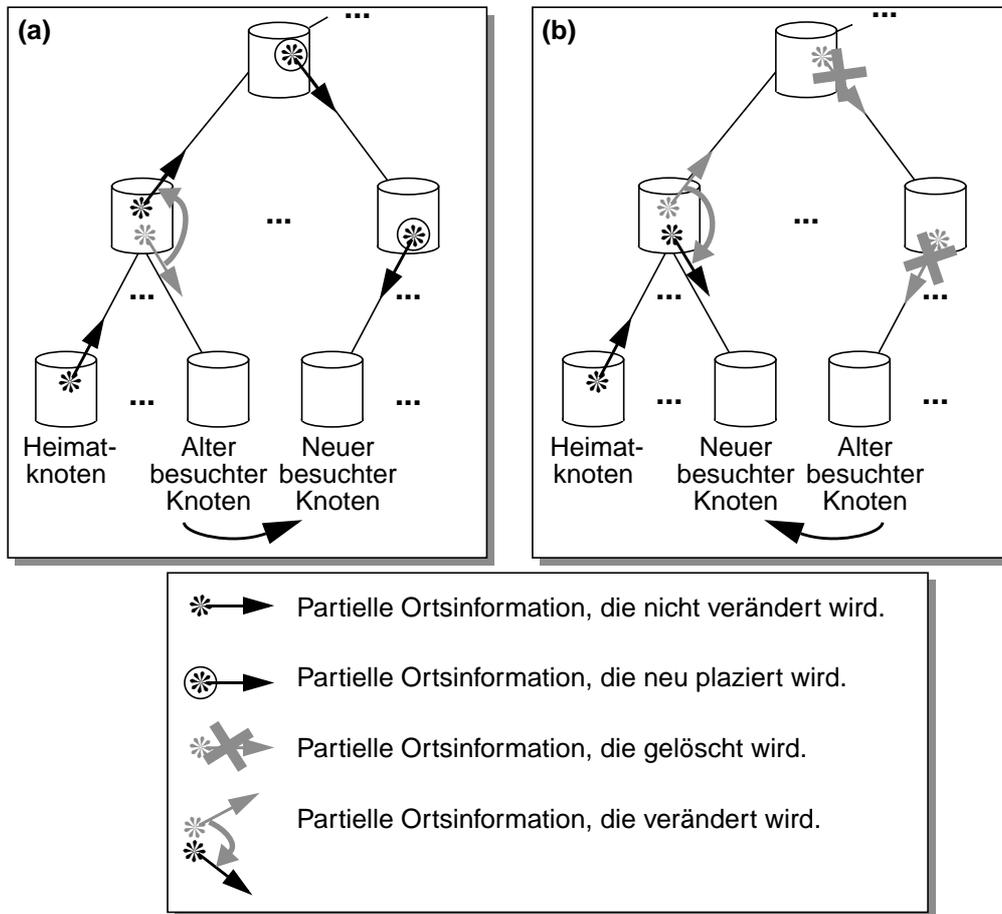


Abbildung 3-7: Zwei Beispiele für die Aktualisierung der partiellen Ortsdaten bei einem Wechsel des besuchten Knotens im hierarchischen Datenbanksystem nach [FSY90] bzw. [Wan93].

Der besondere Vorteile des hier beschriebenen Verfahrens liegt darin, daß der Aufwand für Ortsaktualisierung und Zielsuche im wesentlichen unabhängig von der Entfernung des Teilnehmers von seinem Heimatknoten ist. Der Aufwand für Ortsaktualisierung hängt lediglich vom Abstand zwischen altem besuchten Ort und neuem besuchten Ort ab, der Aufwand für die Zielsuche vom Abstand zwischen dem Ursprung der Anfrage und dem aktuell besuchten Ort. Zielsuche und die Signalisierung zur Ortsaktualisierung müssen entlang der Datenbankhierarchie erfolgen, während andere Verfahren – wie zum Beispiel die GSM-Mobilitätsverwaltung – bei nicht-hierarchischen Signalisieretzstrukturen unter Umständen günstigere Signalisierungswege wählen können und damit gegebenenfalls einen gewissen Vorteil haben. Die Kopplung von Wegführung und Zielsuche muß als Nachteil des Verfahrens betrachtet werden, da die Nutzverkehrsführung nicht mehr unabhängig und flexibel erfolgen kann. Daß diese Kopplung in einem System mit hierarchischer Datenhaltung nicht notwendigerweise erfolgen muß, zeigt der nachfolgende Abschnitt.

3.5.3 Hierarchisches Datenverzeichnis in MONET

Im europäischen Forschungsprojekt MONET¹³ wurden Konzepte und Architekturvorschläge für Mobilkommunikationssysteme der dritten Generation – insbesondere für UMTS – entwickelt. Für die Datenhaltung von UMTS wurde eine sogenannte *Distributed Database* (DDB) vorgeschlagen und Realisierungsformen derselben diskutiert [MON94]. In [Law95] werden einige dieser Realisierungsformen untersucht und auf der Basis einer entsprechenden Modellierung des Teilnehmerverhaltens hinsichtlich Signalisierlast und Belastung der Datenbankknoten durch Datenbankoperationen verglichen.

Das Konzept der DDB sieht zwei Arten von Datenbankknoten vor: *Datenknoten* und *Verzeichnisknoten*. Datenknoten speichern die eigentlichen Daten. Das verteilte Datenverzeichnis, welches aus Verzeichnisknoten in Form eines Baumes und aus Datenknoten, den Blättern des Baumes, gebildet wird, dient zum Ausfindigmachen der Daten. Wegführung und Zielsuche sind hierbei entkoppelt, da vor dem Verbindungsaufbau entsprechende Daten zur Wegführung von der DDB erfragt werden. Anders als bei den Vorschlägen für universelle Mobilität (s. Kapitel 3.5.2) ist hier die Knotenhierarchie auf ein sogenanntes UMTS-Teilgebiet¹⁴ (*domain*) beschränkt. Anfragen aus einem Teilgebiet werden an die Datenknoten gestellt und gegebenenfalls an Wurzelknoten fremder Teilgebiete via den eigenen Wurzelknoten weitergereicht. Der Vorteil dieses Verfahrens besteht darin, daß jedes UMTS-Teilgebiet unabhängig von den anderen Teilgebieten organisiert und verwaltet werden kann.

Für unterschiedliche Kategorien von Daten lassen sich unterschiedliche Platzierungsverfahren realisieren. Für ein bestimmtes Datum innerhalb eines UMTS-Teilgebietes wird dabei im allgemeinen zwischen einem sogenannten *resident node* („Heimatknoten des Datums“), in dem das Datum längerfristig plaziert wird oder zumindest eine Verweiszeiger auf den Ort des Datums existiert, und einem sogenannten *visited node* („besuchter Knoten des Datums“), in dem das Datum oder eine Kopie des Datums temporär plaziert ist, unterschieden. In [MON94] werden drei Basis-Alternativen für die Platzierung eines Datums präsentiert:

- ❑ nur im Heimatknoten,
- ❑ nur im besuchten Knoten,
- ❑ das Primärdatum im Heimatknoten und eine Kopie im besuchten Knoten¹⁵.

Die Platzierung eines Datums wird im verteilten Datenverzeichnis verzeichnet. Ein Eintrag in einem Verzeichnisknoten enthält einen Verweiszeiger auf das Primärdatum oder auf einen Heimatknoten, in dem sich lediglich ein Verweiszeiger auf das Primärdatum befindet, sowie unter

13. Mobile Networks, R2066, RACE II, 3. Forschungsrahmenprogramm der Europäischen Union, 1990-1994

14. Ein UMTS-Teilgebiet kann beispielsweise das Netz eines einzelnen Netzbetreibers sein. Jeder Teilnehmer ist einem UMTS-Teilgebiet zugeordnet.

15. Als Erweiterung dieser Alternative wird vorgeschlagen, mehrere Kopien an verschiedenen Stellen zu plazieren.

Umständen auch Verweiszeiger auf Kopien des Datums. Daneben können jedoch auch Verweiszeiger auf Kopien im Datenknoten des Primärdatums gespeichert sein. Zwei Arten von Verweiszeigern können unterschieden werden:

- direkte Verweiszeiger, die direkt auf einen Datenbankknoten zeigen, und
- indirekte Verweiszeiger, die auf andere Verzeichnisknoten zeigen, in denen sich wiederum Einträge bezüglich des betreffenden Datums befinden.

Je nachdem, ob die Platzierung von Daten stabil bleibt oder sich dynamisch ändert, und je nach Verfahren zum Ausfindigmachen und Aktualisierung der Daten, bieten sich verschiedene Möglichkeiten, entsprechende Einträge im Datenverzeichnis zu platzieren. Ändert sich beispielsweise die Platzierung des Primärdatums nicht, so bietet es sich an, direkte (statische) Zeiger auf das Primärdatum zu verwenden. Zumindest im Wurzelknoten sollte sich ein Eintrag mit solch einem statischen Zeiger befinden, um Anfragen von fremden UMTS-Teilgebieten direkt weiterreichen zu können und für interne Anfragen zumindest einen Verweiszeiger zu haben, der leicht gefunden werden kann. Der Ort des Primärdatums (und damit in der Regel auch der Heimatknoten) kann um so schneller ermittelt werden, in je mehr weiteren Verzeichnisknoten sich entsprechende Einträge befinden. Kann sich dagegen die Platzierung eines Primärdatums oder einer Kopie ändern, so kann sich auch der Inhalt der Einträge mit Verweiszeigern und gegebenenfalls auch ihre Platzierung dynamisch ändern. Hier bietet sich unter Umständen die Verwendung indirekter Zeiger an.

Abbildung 3-8 zeigt die logische Struktur der MONET-DDB sowie ein Beispiel, bei dem das Datum nur im besuchten Knoten gehalten wird und bei dem mehrere direkte (statische) Zeiger auf seinen Heimatknoten im Verzeichnissystem verteilt sind. Daneben werden indirekte Zeiger im Heimatknoten und im Verzeichnissystem verwendet, die eine Kette vom Heimatknoten zum besuchten Knoten bilden.

Anfragen an das Datenbanksystem werden mit Hilfe der Zeiger in den Verzeichnisknoten und möglichen Verweiszeigern auf besuchte Knoten im Heimatknoten zwischen Knoten der DDB weitergereicht. Alle bereits beschriebenen Weiterreichungstechniken (*Verweisen*, *Verketteten*, *Abgeben* und *Multicasting*) sind zulässig. Generell gilt, daß eine Anfrage, für die kein Eintrag in einem DDB-Knoten gefunden wird, zum nächsten Verzeichnisknoten in Richtung Baumwurzel weitergereicht wird. Wird im Verzeichnissystem ein Eintrag gefunden, der sowohl auf ein Primärdatum als auch auf eine Kopie verweist bzw. sowohl auf einen Heimatknoten als auch auf einen besuchten Knoten, so besteht die Option, zwischen den Verweiszeiger zu wählen. Ob solch eine Wahl stattfindet, welche Kriterien dabei verwendet werden und welche Weiterreichungstechnik benutzt wird, definiert die Methode zum Ausfindigmachen von Daten. In [MON94] wird eine Reihe solcher Methoden diskutiert, die aus Kombinationen verschiedener Weiterreichungstechniken mit drei einfachen Suchregeln (R, V und RV) besteht. Es wird dabei beispielhaft davon ausgegangen, daß ein Datum sich ausschließlich im besuchten Knoten befindet. Für die Suchregel R wird die Anfrage zum Heimatknoten geführt und von dort mit Hilfe eines Verweiszeigers an den besuchten Knoten weitergereicht. Einträge im Verzeichnis-

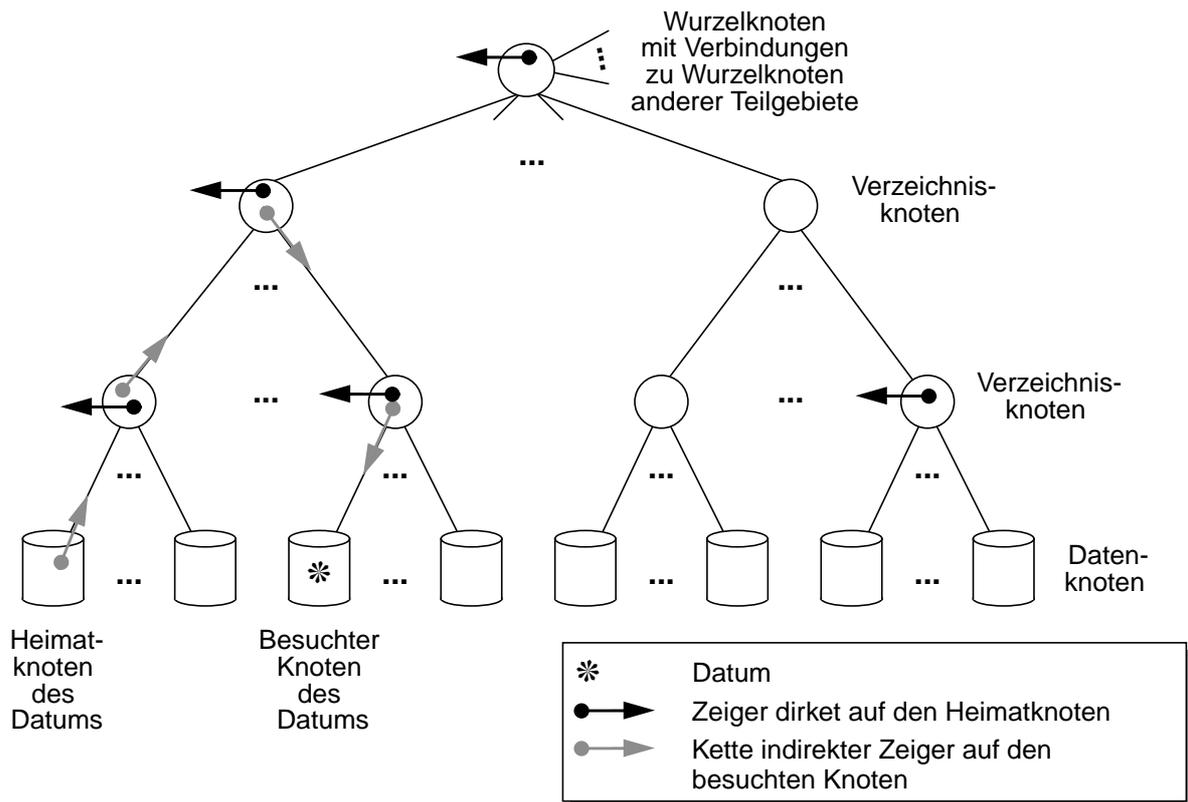


Abbildung 3-8: Logische Struktur der hierarchischen MONET-DDB (eines UMTS-Teilgebietes), in die exemplarisch die Plazierung eines beliebigen Datums sowie beispielhafte Einträge im Verzeichnissystem und im Heimatknoten des Datums eingezeichnet sind.

system, die auf den besuchten Knoten verweisen, sind nicht notwendig. Bei der Suchregel V wird die Anfrage direkt zum besuchten Knoten geführt. Dazu müssen geeignete dynamische Verweiszeiger im Verzeichnissystem existieren. Die Suchregel RV besagt, daß eine Anfrage solange nach den Regeln der Suchregel R geführt wird, bis ein Verweis auf den besuchten Knoten gefunden wird.

Die Aktualisierung der Daten wird durch entsprechende Datenbankoperationen getrieben. Es ist Aufgabe desjenigen Datenknotens, der das Primärdatum hält, Kopien zu aktualisieren. Ändert sich die Plazierung eines Datums, so ist es Aufgabe des Verzeichnissystems, die entsprechenden Einträge zu aktualisieren. Die dazu gewählte Vorgehensweise hängt von der Art und der Plazierung der Einträge im Verzeichnissystem ab. Geht man davon aus, daß für ein Datum ein Heimatknoten bekannt ist, so kann man die Plazierung der Verzeichniseinträge für ein Datum im besuchten Knoten relativ zum Heimatknoten definieren. In [MON94] sind vier mögliche Varianten aufgeführt (siehe Abbildung 3-9):

- a) Einträge in allen Knoten im Pfad vom Wurzelknoten zum besuchten Knoten.
- b) Einträge in allen Knoten im Pfad vom Heimatknoten zum besuchten Knoten.
- c) Nur ein Eintrag. Und zwar in dem Verzeichnisknoten im Pfad vom Heimatknoten zum besuchten Knoten, der dem Wurzelknoten am nächsten ist (Brückenknoten).

- d) Zwei Einträge. Jeweils im Knoten unterhalb des Brückenknotens in Richtung von Heimatknoten bzw. besuchtem Knoten.

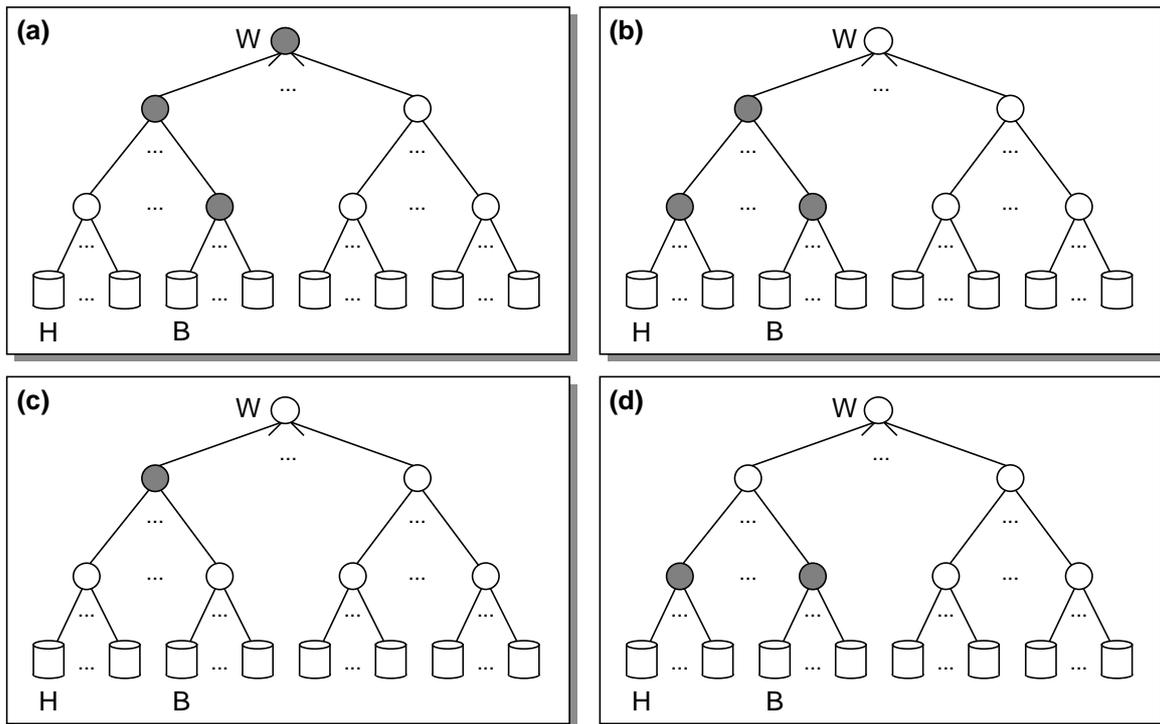


Abbildung 3-9: Vier mögliche Varianten zur Platzierung von Verweiszeigern (●) zu Daten in einem besuchten Knoten (B) in Relation zum Heimatknoten (H) und zum Wurzelknoten (W).

Für die Varianten *a* und *b* bietet sich die Verwendung von indirekten Zeigern an. Für die Varianten *c* und *d* kommen nur direkte Zeiger in Frage. Für die numerischen Untersuchungen beschränkt sich [Law95] auf die Variante *a*.

Geht man davon aus, daß die Ortsinformation, ähnlich wie in den zuvor beschriebenen Systemen, am besuchten Ort eines Teilnehmers vergeben wird, so kann man für die Platzierung der Ortsinformation davon ausgehen, daß sie nur einfach vorhanden ist und im besuchten Knoten gehalten wird. Damit ist hierfür der Fall gegeben, der bei der Vorstellung der drei Suchtechniken *R*, *V* und *RV* beschrieben wurde. Für die Ortsinformation bieten sich alle dargestellten Methoden und Varianten zum Ausfindigmachen von Daten, zur Verweiszeigerplatzierung und zur Datenaktualisierung an.

Befindet sich ein Teilnehmer in einem fremden UMTS-Teilgebiet, so hängt es von den dort vorgesehenen Datenverwaltungsstrategien ab, welche Einträge im Verzeichnissystem vorgenommen werden und welche Arten von Datenknoten (*Heimatknoten*, *besuchter Knoten*) ihm zugewiesen werden. Kann eine Anfrage in einem UMTS-Teilgebiet nicht beantwortet werden, so wird sie vom Wurzelknoten an das Heimat-Teilgebiet des Teilnehmers weitergereicht. Befindet sich das gesuchte Datum nicht dort, so existiert ein Verweiszeiger zum besuchten UMTS-Teilgebiet. Typischerweise befindet sich der Verweiszeiger im Heimatknoten, der in Abhängigkeit von der Datenplatzierungs- und -ausfindigmachungsstrategie (z.B. für die Suche nach der *V*-Suchtechnik) auch die Rolle eines *besuchten Knotens* spielen kann.¹⁶ Wie bei GSM

– und anders als bei dem in Kapitel 3.5.2 beschriebene System – laufen Ortsanfragen aus den meisten fremden UMTS-Teilgebieten über den Heimatknoten im Heimat-Teilgebiet eines Teilnehmers, der sich nicht in seinem Heimat-Teilgebiet befindet.

Die Datenhaltung der MONET-DDB ist so flexibel, daß sie an spezifische Netzstrukturen und an das Bewegungs- und Rufverhalten der Teilnehmer angepaßt werden kann. Innerhalb eines UMTS-Teilgebiets skaliert sie in der Regel sehr gut. Durch die organisatorische Trennung der UMTS-Teilgebiete kann es jedoch zu einer starken Beanspruchung der Wurzelknoten kommen, wenn Teilnehmer häufig aus anderen Teilgebieten gerufen werden.

Die Verfolgung eines Teilnehmers ist in solch einem System in der Regel recht einfach, da lediglich eine Anfrage an die DDB gestellt werden muß, um dem aktuellen Aufenthaltsort eines Teilnehmers (d. h. den aktuell besuchten Knoten) zu ermitteln. Wird in Heimatknoten ein Verweiszeiger auf den aktuell besuchten Knoten gehalten, so könnte dort – wie auch im GSM-HLR – sehr einfach ein Bewegungsprofil erstellt werden.

3.5.4 *Mobile IP mit Route Optimization*

Die Datenhaltung für *Mobile IP* [Per96] ist sehr einfach. Die Ortsinformation (*care-of address*) befindet sich im *Home Agent*. Alle IP-Pakete werden dorthin geführt und von dort weitergeleitet. Aktualisierung erfolgt durch den mobilen Rechner bei Ablauf eines Timers oder bei Ortswechsel. Bleibt eine Aktualisierung nach Ablauf der Lebensdauer aus, so wird die Ortsinformation im *Home Agent* verworfen und der mobile Rechner wird unerreichbar. Eine explizite Deregistrierung beim zuletzt besuchten *Foreign Agent* ist nicht vorgesehen.

Die nicht optimale Verkehrsführung (Dreiecks-Routing) und die unter Umständen massiven Paketverluste beim Wechsel des *Foreign Agent* sind Nachteile von *Mobile IP*, die mit Hilfe des im *Internet-Draft* „*Route Optimization in Mobile IP*“ [PJ97] vorgeschlagenen Protokolls beseitigt werden sollen. Es kommt dabei zu einer Vervielfachung der Ortsinformation. Kopien werden an die Kommunikationspartner eines mobilen Rechners geschickt. Diese können dann die IP-Pakete, die für den mobilen Rechner bestimmt sind, direkt zur *care-of address* „tunneln“, ohne daß diese zuerst zum *Home Agent* gelangen müssen. Die Kommunikationspartner eines mobilen Rechners erhalten die *care-of address* entweder auf Anfrage (mit Hilfe einer *Binding Request Message*) oder dadurch, daß sie IP-Pakete zur permanenten Adresse des mobilen Rechners schicken und der *Home Agent*, der diese Pakete erhält, selbsttätig mit einer *Binding Update Message* den Absender der IP-Pakete über die *care-of address* informiert. Dazu muß die Authentizität von Meldungen zwischen dem Kommunikationspartner des mobilen Rechners und dem *Home Agent* des mobilen Rechners – wie in Kapitel 2.5.1 beschrieben – gesichert werden. Die so erhaltene Ortsinformation wird solange von einem

16. Es wäre auch denkbar, den Zeiger in einem besuchten Knoten zu hinterlegen, der nicht der Heimatknoten ist.

Kommunikationspartner verwendet, bis ihre Lebensdauer abgelaufen ist oder bis sie erneuert wird. Es bleibt dem Kommunikationspartner überlassen, ob er vor Ablauf der Lebensdauer der Ortsinformation eine erneute Anfrage an den *Home Agent* stellt.

Bei einer Änderung der Ortsinformation ist es nicht die Aufgabe des *Home Agent*, unmittelbar Ortsinformation bei den Kommunikationspartnern des mobilen Rechners zu aktualisieren. Erst wenn aufgrund veralteter Ortsinformation IP-Pakete zu einem *Foreign Agent* gelangen, bei dem der adressierte mobile Rechner nicht mehr registriert ist, stößt der *Foreign Agent* beim *Home Agent* mit Hilfe einer *Binding Warning Message* eine Ortsinformationsaktualisierung an. Da jedoch im ursprünglichen Mobile-IP-Protokoll ein *Foreign Agent* nicht unmittelbar feststellen kann, ob sich ein mobiler Rechner bewegt und bei einem anderen *Foreign Agent* registriert hat, wird in [PJ97] eine explizite Deregistrierung vorgesehen. Dazu teilt der neue mobile Rechner dem zuletzt besuchten *Foreign Agent* direkt seine neue *care-of address* mit (*Previous Foreign Agent Notification*). Dies hat dazu noch den Vorteil, daß der zuletzt besuchte *Foreign Agent* IP-Pakete, die aufgrund veralteter Ortsinformation zu ihm gelangen, zur neuen *care-of address* weiterleiten kann. Dadurch kann unter Umständen ein Großteil der Paketverluste beim Wechsel des *Foreign Agent* vermieden werden, was solch eine Nutzverkehrsführungsänderung nahtloser gestaltet („*smooth handoff*“). Selbst für den Fall, daß ein Kommunikationspartner *Mobile IP* nicht unterstützt, können sich Vorteile aus der expliziten Deregistrierung und dem Weiterleiten der IP-Pakete ergeben. Der vorherige *Foreign Agent* kann meist relativ schnell über einen Ortswechsel informiert werden. Er kann dann die Pakete weiterleiten, die der *Home Agent* verschickt, solange ihn die neue Registrierungsmeldung noch nicht erreicht hat.

Abbildung 3-10 stellt die Datenhaltung bei *Mobile IP* mit *Route Optimization* dar. Hier liegt nun ein Fall vor, bei dem die Mobilitätsverwaltung mit Hilfe sogenannter *Cache*-Kopien arbeitet, deren Übereinstimmung mit dem Primärdatum nicht gewährleistet ist. Pakete werden unter Umständen an den „falschen“ *Foreign Agent* geschickt. Durch das Weiterleiten der Pakete zum neuen *Foreign Agent* werden jedoch die negativen Folgen der veralteten Information gemildert. Anstelle aufwendiger Mechanismen zur Konsistenzhaltung der Daten wird ein Verfahren verwendet, das erst bei Bedarf – und unter Inkaufnahme von erhöhtem Aufwand zur Nutzverkehrsführung – die Ortsinformationskopien aktualisiert.

Anders als bei klassischen mobilen Telekommunikationssystemen, wo die aktuelle Aufenthaltsrufnummer eines Teilnehmers den rufenden Endgeräten verborgen bleibt, wird bei *Mobile IP* mit *Route Optimization* die *care-of address* an den Kommunikationspartner weitergegeben. Somit kann im Prinzip jeder Kommunikationspartner Informationen über den aktuellen Aufenthaltsort eines mobilen Rechners erhalten.

So wie bei GSM Signalisierung mit dem HLR stattfindet, so findet bei *Mobile IP* mit *Route Optimization* Signalisierung mit dem *Home Agent* statt. Anders als beim GSM-HLR kann der *Home Agent* jedoch nicht an einer beliebigen Stelle im Netz plaziert werden, sondern muß unter der permanenten IP-Adresse des mobilen Rechners erreichbar sein. Die Möglichkeit, durch günstige Platzierung Signalisierungskosten und -verzögerung zu reduzieren, entfällt damit.

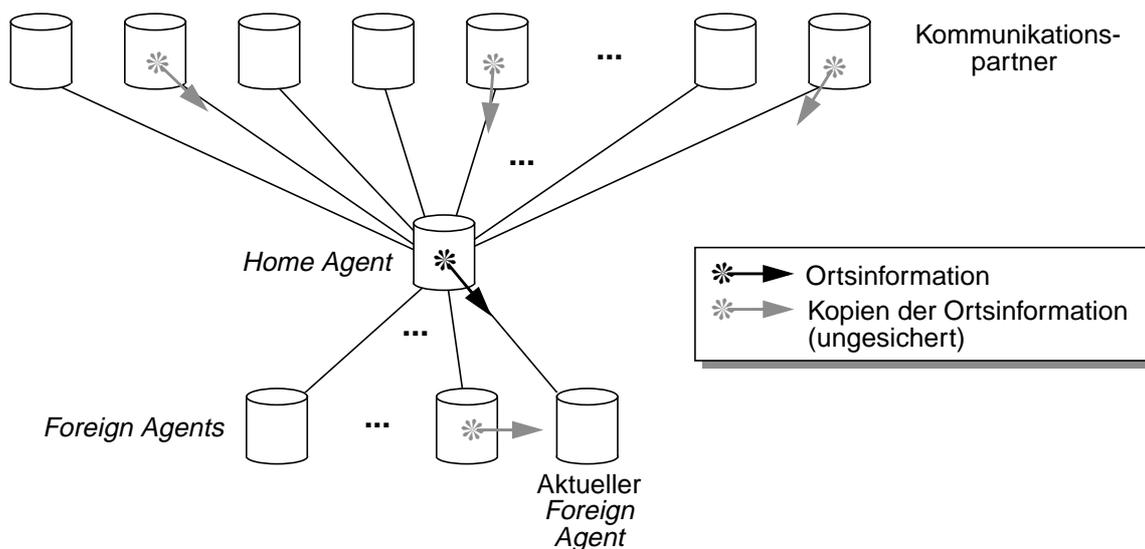


Abbildung 3-10: Datenhaltung bei *Mobile IP* mit *Route Optimization*.

Auch der Einsatz zuverlässiger Netzknoten, die für sehr viele Teilnehmer die Funktion eines *Home Agent* übernehmen, wird dadurch erschwert. Da die Funktionsfähigkeit von *Mobile IP* jedoch von der Verfügbarkeit des *Home Agent* abhängt, ist es wichtig, daß dieser gut erreichbar ist und zuverlässig in Betrieb bleibt. Im Vergleich zu *Mobile IP* haben die Konzepte, die in klassischen mobilen Telekommunikationsnetzen Anwendung finden, diesbezüglich in der Regel deutliche Vorteile, da einige wenige, gut platzierte und zuverlässige Netzknoten die Mobilitätsunterstützung sicherstellen können.

3.5.5 Die *Datacycle Architecture*

Ein radikal anderer Ansatz zur Datenhaltung für IN-Dienste und personenbezogene Mobilität wird in [BGH90] vorgeschlagen. Während einige Ansätze versuchen, Daten dadurch schnell verfügbar zu machen, daß sie lokal gehalten bzw. gespeichert werden, werden Daten bei der sogenannten *Datacycle Architecture* dadurch schnell verfügbar gemacht, daß sie sehr häufig auf einem *Broadcast*-Medium (zum Lesen) zur Verfügung gestellt werden. Ein Primärdatensatz wird logisch zentral gehalten. Alle Änderungen werden direkt an ihm vorgenommen. Der Inhalt des Primärdatensatzes wird periodisch auf einem *Broadcast*-Medium ausgesendet. Soll auf diesen Datensatz lesend zugegriffen werden, so wird das *Broadcast*-Medium von einem sogenannten Datenfilter in der anfragenden Instanz solange abgehört, bis der Datensatz gesendet wurde. Obwohl keine Kopien von Primärdatensatz erstellt werden, kann man das Verfahren im Grunde dennoch als ein *primary shadowing* Verfahren mit einem sehr häufigen periodischen Abgleich klassifizieren.

Um kurze Zugriffszeiten zu ermöglichen, wird in [BGH90] vorgeschlagen, so viel Bandbreite zur Verfügung zu stellen, daß der gesamte Datenbestand in einer Zyklusdauer von deutlich unter einer Sekunde gesendet werden kann. Dabei wird eine logische und physikalische Aufteilung des *Broadcast*-Mediums in Datenkanäle notwendig. Geht man davon aus, daß auf einem Datenkanal mit 2 Gbit/s gesendet wird und daß ein Datenfilter diese Datenmenge auch

verarbeiten kann, so ergibt sich, daß – bei einer angestrebten Zykluszeit von circa 0,4 Sekunden – für je 100 MByte Teilnehmerdaten ein Datenkanal verwendet werden muß. Damit bei einer lesenden Anfrage nicht alle Datenkanäle (unter Belegung von je einem Datenfilter) abgehört werden müssen, ist es sinnvoll, Informationen darüber zu haben, welche Daten auf welchem Datenkanal gesendet werden.¹⁷ Abbildung 3-11 zeigt das Prinzip der *Datacycle Architecture*. Beispielhaft sind zwei Datenkanäle eingezeichnet, die von einem einzelnen Datenbankknoten gespeist werden. In einem System mit mehreren Datenbankknoten kann man gezielt nur die Datenkanäle abhören, die von dem für den Teilnehmer zuständigen Knoten gespeist werden.

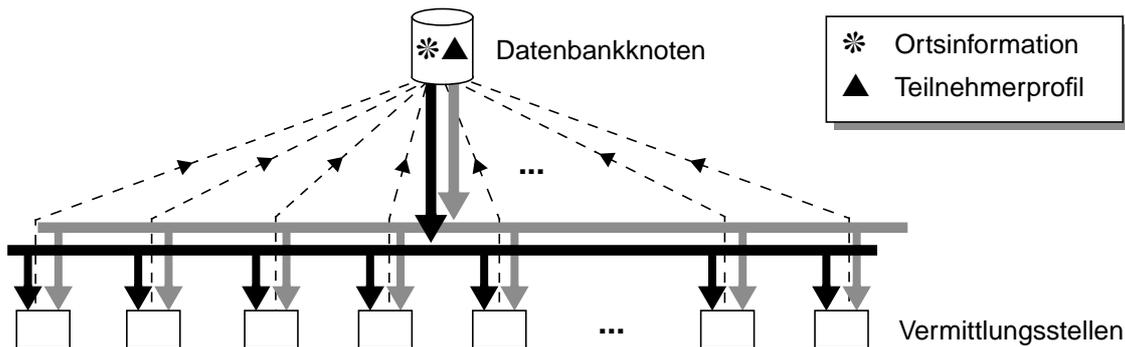


Abbildung 3-11: Prinzip der *Datacycle Architecture*, dargestellt mit zwei *Broadcast*-Datenkanälen, die von einem einzelnen Datenbankknoten gespeist werden.

3.5.6 Mobile PNNI

In der *Wireless ATM Working Group* des *ATM Forum* wurde einige Zeit lang ein Vorschlag [VD96a, VD96b] diskutiert, der die Mobilitätsverwaltung für mobile Endsysteme in das Wegführungsprotokoll PNNI integriert. Dieser Ansatz wird als *Mobile PNNI* bezeichnet. Er ist einer der wenigen Ansätze, bei denen die Wegführung auf Basis der permanenten Adresse in der Vermittlungsschicht erfolgt. Mit Hilfe von *PNNI Topology State Packets* werden die Wegführungstabellen so beeinflusst, daß Wegführung direkt zum aktuellen Netzzugangspunkt eines Endsystems erfolgt. Ein großer Nachteil dieses Verfahrens ist der hohe Aufwand, der zur Verbreitung und Speicherung der Erreichbarkeitsinformation je mobilem Endsystem getrieben werden muß. Darüberhinaus wird unter Umständen die Leistungsfähigkeit der Wegführung zu festen Endsystemen beeinträchtigt, da größere Verkehrslenkungstabellen verwaltet werden müssen. Um diese Nachteile zu begrenzen, ist in [VD96a] vorgesehen, die Verbreitung der Erreichbarkeitsinformation auf einen gewissen Bereich¹⁸ in der Umgebung des aktuellen Aufenthaltsorts eines mobilen Endsystems zu beschränken. Dadurch werden nur noch Verbindungen, deren Ursprung sich in diesem Bereich befindet, auf direktem Weg zum Endsystem

17. In [BGH90] wird ein hierarchisches Verfahren angedeutet, bei dem auch diese Information über das *Broadcast*-Medium verfügbar gemacht wird.

18. PNNI bildet Netzknoten auf eine hierarchische Struktur ab. Ein Bereich, in dem Erreichbarkeitsinformation verbreitet wird, kann durch den Abstand zwischen zwei Knoten innerhalb dieser Hierarchie definiert werden.

geführt. Alle anderen Verbindungen werden – wie bei *Mobile IP* – zuerst zur Heimatvermittlungsstelle geführt und von dort zum aktuellen Aufenthaltsort weitergeleitet. Dazu ist es erforderlich, die Heimatvermittlungsstelle über jeden Aufenthaltsortwechsel zu informieren und die Signalisierungsprotokolle so zu verwenden, daß ein Bezeichner für das mobile Endsystem bei der Verbindungsweiterleitung in der Verbindungsaufbau-Meldung mitgeführt wird [VD96b]. Wie bei *Mobile IP* kommt es hierbei zu nicht optimaler Verbindungsführung. In [VD96b] werden deshalb verschiedene Möglichkeiten angesprochen, wie die Verbindungsführung nachträglich optimiert werden könnte. Der Vorteil von *Mobile PNNI* besteht darin, daß bei der Adreßvergabe keine Unterscheidung zwischen mobilen und festen Endsystemen getroffen werden muß. Durch den hohen Aufwand zur Ortsaktualisierung und zur nachträglichen Optimierung der Verbindungsführung sowie wegen der Nachteile aufgrund der Vergrößerung der Verkehrlenkungstabellen, ist das Verfahren jedoch nur für eine geringe Zahl von mobilen Endsystemen geeignet, die sich nur relativ selten bewegen. In der *Wireless ATM Working Group* wird für *Wireless ATM* inzwischen ein Verfahren zur Mobilitätsverwaltung vorgeschlagen, das ähnlich wie bei anderen Systemen nach der Befragung einer Datenbank (*Location Server*) eine Adreßumsetzung vornimmt [ATMF97, BG98].

3.5.7 Erweiterungen der vorgestellten Verfahren

Die bisher in diesem Unterkapitel vorgestellten Systeme und Verfahren zur Mobilitätsverwaltung stellen nur eine sehr begrenzte Auswahl aus der Vielzahl von veröffentlichten Vorschlägen dar. Bei der Auswahl dieser Beispiele wurde darauf geachtet, daß sie in ihrer Gesamtheit das Spektrum der Möglichkeiten ausreichend abdecken und daß sie entweder von besonderer technischer Relevanz oder von besonderer Bedeutung für andere Publikationen und Vorschläge sind. Um den letzten Punkt zu belegen, werden in diesem Abschnitt einige Publikationen vorgestellt, die Variationen oder Erweiterungen der vorgestellten Systeme beinhalten. Einige Vorschläge, die das bestehende GSM-System modifizieren, werden jedoch erst in Kapitel 4.2 diskutiert.

In [vSHT96] wird zur Lokalisierung virtueller mobiler Objekte, wie zum Beispiel von Dokumenten im *World Wide Web* oder von *computational objects* in TINA, eine verteilte Datenbank vorgeschlagen. Anders als bei einem Mobilkommunikationssystem kann es von solch einem Objekt mehrere identische und gleichwertige Kopien geben, so daß das Objekt an mehreren Stellen erreichbar sein kann. Vergleichbar mit der MONET-DDB bilden bei diesem Vorschlag Verzeichniszeiger eine Kette von der Wurzel bis zu dem Blatt der Datenbank, an dem die Abbildung von einem Objektbezeichner auf eine aktuelle „Kontaktadresse“ (Adreßumsetzung) vorgenommen werden kann. Entsprechende Abzweigungen von der Kette verweisen auf alternative Datenbankblätter zur Adreßumsetzung. In einer Erweiterung des Verfahrens ist vorgesehen, durch geeignete Verfahren die Häufigkeit von Veränderungen im Verzeichnissystem zu minimieren und, anders als bei der MONET-DDB, die Adreßumsetzung auch in Knoten höherer Ebenen der Datenbankhierarchie vorzunehmen. Schließlich werden zur Beschleunigung der Suche *Cache*-Kopien von Zeigern auf Adreßumsetzungsknoten vorgeschlagen.

Neben *Mobile PNNI* wird in [VD96a] auch ein Verfahren mit einem Verzeichnissystem, das aus mehreren Teilbäumen besteht, vorgeschlagen. Die Hierarchie der Verzeichnisknoten orientiert sich an der Hierarchie der PNNI-Wegführung. In einem besuchten Teilbaum wird eine Kette von Zeigern von der Wurzel zum besuchten Ort gehalten. In einem Heimatverzeichnisknoten wird auf den besuchten Teilbaum verwiesen. (Siehe Abbildung 3-12). Die Suche nach einem Teilnehmer beginnt beim anfragenden Blatt. Wird die Wurzel des Teilbaums erreicht und kein Hinweis auf den Teilnehmer gefunden, so wird dessen Heimatverzeichnisknoten befragt und die Suche an der Wurzel des besuchten Teilbaumes fortgesetzt. Diese Verzeichnisknotenhierarchie kann auch als ein einziger Baum betrachtet werden, bei dem ein teilnehmerindividueller logischer Wurzelknoten existiert (der Heimatverzeichnisknoten), unterhalb dessen sich alle Teilbäume befinden.

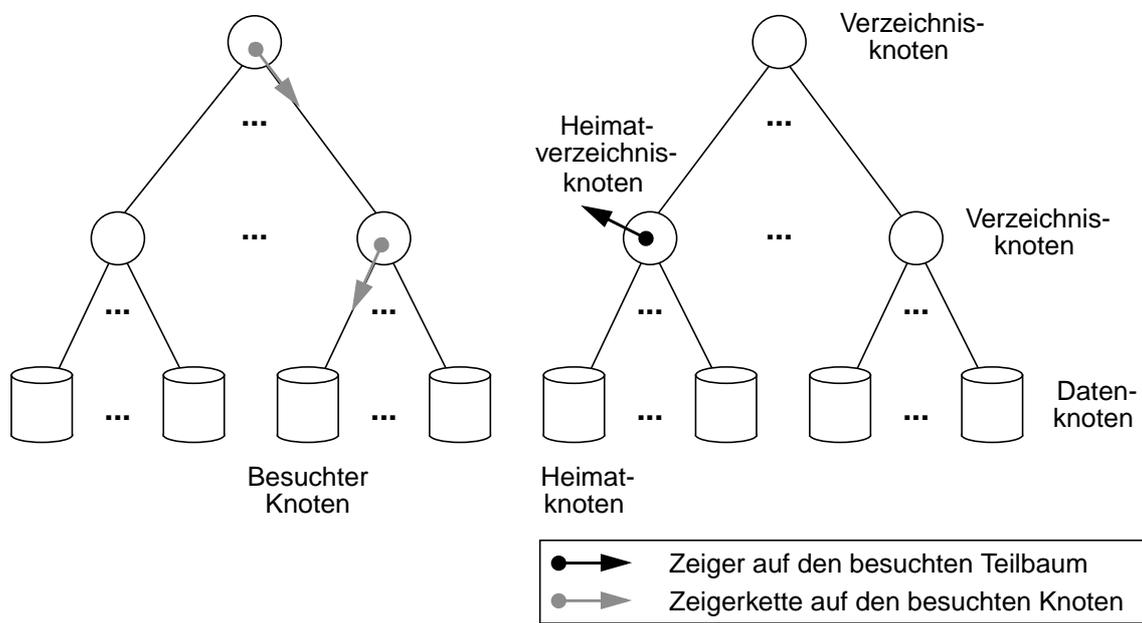


Abbildung 3-12: Logische Struktur der Verzeichnissysteme aus [VD96a].

In [LJCW96] wird eine Modifizierung des Verfahrens nach [Wan93] vorgeschlagen. Anstelle der Kette von Verweiszeigern vom Heimatort zum besuchten Ort wird eine Kette verwendet, die lediglich zwischen Knoten oberhalb einer gewissen Hierarchieebene verläuft. Solange ein Teilnehmer den Abdeckungsbereich eines Knotens dieser Hierarchieebene nicht verläßt, muß die Verweiszeigerkette nicht modifiziert werden, was Signalisierverkehr sparen kann. Andererseits muß bei der Teilnehmersuche der aktuelle Aufenthaltsort durch Ausrufen (*Paging*) in allen Blättern unterhalb des besuchten Knotens der gewählten höheren Hierarchieebene ermittelt werden, was aufwendige Signalisierung zur Folge haben kann. In [FHP95] wird vorgeschlagen, die geeignete Hierarchieebene, bis zu der keine Information über den Aufenthaltsort eines Teilnehmers gehalten wird, – und damit die Zahl der beteiligten Knoten beim Ausrufen – dynamisch an das Teilnehmergehalten anzupassen.

Kapitel 4

Vorschläge zur Weiterentwicklung von Verfahren der Mobilitätsverwaltung

In den letzten beiden Kapiteln wurde die Problematik der Mobilitätsunterstützung in Kommunikationsnetzen systematisiert. Es wurde dargelegt, in welcher Weise das Phänomen Mobilität zu Tage tritt und mit welchen Mitteln es unterstützt werden kann. Die grundlegenden unterstützenden Netztechniken wurden beschrieben und die Aufgaben der Mobilitätsverwaltung definiert. Verschiedene Möglichkeiten im Zusammenspiel von Wegführung und Zielsuche sowie in der Datenhaltung wurden systematisch aufgezeigt. Der Wert dieser Ausführungen soll nun in diesem Kapitel dadurch belegt werden, daß auf ihrer Basis einige Vorschläge zur Weiterentwicklung von mobilitätsunterstützenden Systemen und Verfahren erarbeitet werden.

Die Vorschläge sind in drei Gruppen gegliedert: Vorschläge, welche optimale Nutzverkehrsführung ermöglichen (Kapitel 4.1), Vorschläge, welche die GSM-Mobilitätsverwaltung fortentwickeln (Kapitel 4.2), und Vorschläge, welche Mobilität über verschiedene Netztechniken hinweg unterstützen (Kapitel 4.3). Besonders hervorzuheben sind dabei die drei folgenden neuen Verfahren:

- ❑ Das Senden von frühen *Binding Updates* für *Mobile IP* (Kapitel 4.1.3),
- ❑ die Einführung von *Proxy HLRs* in GSM (Kapitel 4.2.4) und
- ❑ eine *Common Distributed Database* für netzeübergreifende Aufenthaltsortsverwaltung (Kapitel 4.3.2).

In den nachfolgenden Kapiteln werden Methoden und Analysen behandelt, um die Vorschläge hinsichtlich ihrer (quantitativen) Leistungsfähigkeit zu untermauern.

4.1 Unterstützung von optimaler Nutzverkehrsführung

Soll zwischen zwei Endsystemen eine Kommunikationsbeziehung aufgebaut werden, so empfiehlt es sich, die Zielsuche vor der Wegführung und damit unabhängig davon durchzuführen. Im Idealfall sollte Nutzverkehr immer auf dem direkten Weg¹ zwischen zwei Endsystemen

1. Der „direkte Weg“ sei der Weg, den die Wegführung auch zu einem nicht-mobilen Endsystem am selben Netzzugangspunkt wählen würde.

geführt werden. In realen Systemen kann es verschiedene Gründe geben, warum dies nicht geschieht beziehungsweise nicht geschehen kann (vgl. Kapitel 3.2):

- ❑ Netzknoten und Endsysteme sind nicht auf Mobilität eingerichtet, d. h. sie sind nicht in der Lage, eine Zielsuche zu initiieren und eine Adreßumsetzung durchzuführen.
- ❑ Eine Unterscheidung zwischen mobilen und nicht-mobilen Instanzen wird nicht vorgenommen oder ist mit vertretbarem Aufwand nicht möglich.
- ❑ Gesicherte (authentisierte) Ortsinformation kann nicht erlangt werden.
- ❑ Ortsinformation wird wegen ihres vertraulichen Charakters nicht jedem Netzknoten/Endsystem zur Verfügung gestellt.
- ❑ Die Kosten zur Zielsuche vor der Wegführung übersteigen regelmäßig die Kosten, die durch nicht-optimale Verkehrsführung entstehen.
- ❑ Neben den beiden Endsystemen müssen weitere Instanzen an der Kommunikationsbeziehung beteiligt werden, so zum Beispiel Instanzen zur Entgelterfassung, zur Dienststeuerung oder zur Kommunikationsinhaltsüberwachung. Dies kann eine veränderte Nutzverkehrsführung sinnvoll oder notwendig machen.²

Im Folgenden wird eine Reihe von Maßnahmen vorgestellt, welche es ermöglichen, die Häufigkeit nicht-optimaler Wegführung zu verringern.

4.1.1 Mobilitätsunterstützung im PSTN/ISDN

Wie in Kapitel 2.1.3 dargestellt, können in klassischen Telekommunikationsnetzen viele Rufnummern, die nicht unmittelbar zur Verkehrsführung geeignet sind, prinzipiell auch als solche erkannt werden, da der Numerierungsraum mit Hilfe nationaler und internationaler Dienst- und Netzzugangskennzahlen unterteilt ist. Damit ist möglich, die Notwendigkeit einer Rufnummernumsetzung bereits in Vermittlungsstellen eines PSTN oder ISDN zu erkennen. Somit könnte auch hier nach einer entsprechenden Zielsuche eine Rufnummernumsetzung stattfinden. Bei IN-basierter Mobilitätsunterstützung – wie zum Beispiel bei UPT – wird dies sehr einfach mit Hilfe von IN-Funktionalität in den Vermittlungsstellen realisiert. Entsprechende Ziffernfolgen lösen einen IN-Dienstlogikablauf aus, bei dem der Vermittlungsstelle von einer SCF die entsprechende Information zur Rufnummernumsetzung mitgeteilt wird.

Um dies auch für Rufe zu GSM-Teilnehmern zu realisieren, werden hier zwei Möglichkeiten vorgeschlagen:

- ❑ **Vorschlag A:** Die Vermittlungsstellen werden um GMSC-Funktionalität erweitert, d. h. sie können beim HLR eines Teilnehmers dessen Aufenthaltsrufnummer erfragen und verwenden diese zur Verkehrsführung.

2. Dies kann prinzipiell auch bei Kommunikation zwischen nicht-mobilen Endsystemen der Fall sein. Dort besteht jedoch in der Regel eine gewisse Nähe zwischen dem Netzzugangspunkt und den Instanzen zur Entgelterfassung, zur Dienststeuerung und zur Kommunikationsinhaltsüberwachung.

- Vorschlag **B**: Vermittlungsstellen, die bereits IN-Funktionalität (SSF) besitzen, lösen bei Wahl einer GSM-Rufnummer (MSISDN) einen entsprechenden IN-Dienst aus. Die Dienststeuerung übernimmt dann entweder ein um IN-Funktionalität erweitertes HLR (Vorschlag **B1**) oder ein spezieller SCP, der sich gegenüber dem HLR des gerufenen Teilnehmers wie eine *Gateway* MSC verhält (Vorschlag **B2**).

Vorschlag B hat den Vorteil, daß bestehende Vermittlungsstellen lediglich IN-Funktionalität besitzen müssen, die auch für andere Dienste verwendet wird bzw. verwendet werden kann. Sie müssen nicht wie bei Vorschlag A mit GSM-spezifischer Dienstlogik versehen werden. Vorschlag B1 hat den Nachteil, daß dieser Dienst auch im Heimatnetz des gerufenen Teilnehmers unterstützt werden muß.

Für beide Vorschläge gilt, daß zwischen PSTN/ISDN und dem GSM-Netz des gerufenen Teilnehmers TCAP-basierte Signalisierung möglich sein muß. Bleibt das HLR unverändert (Vorschläge A und B2), so muß netzübergreifende MAP-Signalisierung im Rahmen des Dienstmerkmals *Support of Optimal Routing* der GSM-Phase 2+ möglich sein. Dazu gehört auch, daß Instanzen im PSTN/ISDN mit entsprechenden *Global Titles* im Signalisiersystem Nummer 7 adressierbar sind. Agiert das HLR als SCP (Vorschlag B2), so muß Dienststeuerung auf der Basis von netzübergreifender INAP-Signalisierung stattfinden können. In Kapitel 4.3.1 werden weitergehende Konzepte zum Zusammenwirken von IN-Dienstlogik und GSM-Mobilitätsverwaltung vorgestellt.

4.1.2 Auslösen der Zielsuche bei gemeinsamem Numerierungsraum

Teilen sich mobile und nicht-mobile Endsysteme einen gemeinsamen Numerierungsraum, so ist es schwierig, zwischen diesen zu unterscheiden. Häufig wird deshalb in diesem Fall auf das Auslösen einer Zielsuche vor der Wegführung verzichtet. In einigen Systemen wird bei späterem Erkennen der Mobilität eines Endsystems die Nutzverkehrsführung modifiziert. Im ATM-Forum werden z. B. Mechanismen zur Veränderung der Wegführung im Aufbau befindlicher ATM-Verbindungen vorgesehen (*connection redirect*) [BG98] und bei der Realisierung von Netzbetreiberportabilität geographischer Rufnummern in öffentlichen Telekommunikationsnetzen besteht die Möglichkeit, daß ein Verbindungsaufbau von einer Vermittlungsstelle im Netz des ursprünglichen Netzbetreibers³ abgewiesen wird und eine Vermittlungsstelle im Ursprungsnetz nach einer Rufnummernumsetzung den aktuellen Netzzugangspunkt des gerufenen Teilnehmers adressiert. [LS96] unterscheidet hier zwei Varianten:

- Das Netz des ursprünglichen Netzbetreibers stellt dem Ursprungsnetz des Rufs in der Verbindungsabweisungsmeldung die Adresse des neuen Netzzugangspunkts zur Verfügung.

3. Dies ist im allgemeinen der Netzbetreiber eines Landes, der ehemals ein Monopol für den Dienst innehatte. Gegebenenfalls können jedoch auch Teilnehmerrufnummern (siehe Kapitel 2.1.3) in Rufnummernblöcken an verschiedene Netzbetreiber vergeben sein, so daß mit Information darüber der „ursprüngliche Netzbetreiber“ identifiziert ist.

- Im Ursprungsnetz des Rufs muß die Adresse des neuen Netzzugangspunkts ermittelt werden („Zielsuche“).

Wird dagegen in einem Kommunikationssystem generell vor der Wegführung eine Zielsuche ausgelöst, so kann optimale Wegführung ohne eine nachträgliche Modifikation gewährleistet werden. Dies führt im allgemeinen jedoch zu einer Datenbankanfrage für jede Zieladresse und damit häufig auch zu teurer Signalisierung zwischen Netzknoten auch für Rufe zu nicht-mobilen Endsystemen. Um unnötige Signalisierung oder nachträgliche Modifikation der Wegführung zu reduzieren, wird hier ein hybrider Vorschlag präsentiert, bei dem für einen Teil der Endsysteme mit einer begrenzten Sicherheit lokal festgestellt werden kann, ob sie mobil bzw. nicht-mobil sind. Für die verbleibenden Endsysteme kann eine der beiden oben beschriebenen Varianten gewählt werden.

Dieser Vorschlag beruht darauf, daß zu gerufenen Endsystemen häufig bereits in der jüngeren Vergangenheit Rufe aufgebaut wurden und daß die Eigenschaft eines Endsystems *mobil* (bzw. *portiert*) oder *nicht-mobil* (bzw. *nicht-portiert*) zu sein, sich praktisch nur selten ändert. Am Ursprungsort von Rufen wird lokale Information gehalten, die aus einer Liste von Adressen besteht, bei denen in der Vergangenheit eine Adreßumsetzung notwendig war, und/oder aus eine Liste von Adressen, zu denen in der Vergangenheit keine Adreßumsetzung notwendig war. Jede Zieladresse kann damit als *mobil*, *nicht-mobil* oder *unbestimmt* klassifiziert werden. Für mobile Zieladressen wird eine Zielsuche ausgelöst. Dabei kann das hier vorgestellte Verfahren sehr einfach durch lokale *Cache*-Kopien der neuen Zieladressen ergänzt werden. Nicht-mobile Zieladressen werden direkt zur Wegführung verwendet und lösen keine Zielsuche aus. Bei unbestimmten Zieladressen kann je nach System eine Zielsuche ausgelöst oder unterlassen werden. Da die Entscheidung zur Unterlassung der Zielsuche auf veralteter oder fehlender Information beruhen kann, muß eine Rückmeldung erfolgen, falls nachträglich dennoch eine Adreßumsetzung vorgenommen wird. Bei *Mobile IP* mit *Route Optimization* geschieht dies beispielsweise automatisch mit Hilfe einer *Binding Update Message*. In einem öffentlichen Telekommunikationsnetz mit Netzbetreiberportabilität geographischer Rufnummern geschieht dies mit einer Verbindungsabweisungsmeldung, welche dann außerdem noch eine Zielsuche auslösen kann.

4.1.3 Frühes *Binding Update* für *Mobile IP*

Während in klassischen Sprachkommunikationsnetzen die Wegführung des Nutzverkehrs für beide Richtungen gemeinsam erfolgt, wird diese in IP-basierten Netzen für jede Richtung getrennt durchgeführt. Dies hat zur Folge, daß auch bei Rufen von mobilen zu nicht-mobilen Endsystemen eine Adreßumsetzung (und zwar für die Rückrichtung) stattfinden muß. Bei einfachem *Mobile IP* hat dies zur Folge, daß der Nutzverkehr in Rückrichtung über den *Home Agent* geführt wird. Bei *Mobile IP* mit *Route Optimization* wird der Nutzverkehr in Rückrichtung vom Kommunikationspartner zuerst zum *Home Agent* des mobilen Rechners geführt. Erst wenn ein IP-Paket den *Home Agent* erreicht, informiert dieser den Kommunikationspartner mit Hilfe einer *Binding Update Message* über die aktuelle *care-of address*. Hier wird nun ein Vor-

schlag gemacht, bei dem der Kommunikationspartner etwas früher über die aktuelle *care-of address* eines mobilen Rechners informiert wird, der eine Kommunikationsbeziehung zu ihm aufbaut, so daß gegebenenfalls Nutzverkehr schon früher auf dem direkten Weg zwischen nicht-mobilem und mobilem Rechner geführt werden kann. Der Vorschlag bedarf keiner Änderung der Schnittstellen von *Mobile IP* nach [PJ97] und modifiziert lediglich das Verhalten des mobilen Rechners leicht. Es handelt sich um einen zusätzlichen Mechanismus, dessen Fehlen oder Scheitern die Funktionsweise von *Mobile IP* mit *Route Optimization* nicht beeinträchtigt.

Häufig ist ein mobiler Rechner ein Endgerät eines Internet-Nutzers, der zum Beispiel von (nicht-mobilen) Servern Informationen anfordert. Die Initiative zum Aufbau einer Kommunikationsbeziehung geht dabei oft vom mobilen Rechner aus. Es wird deshalb vorgeschlagen, daß ein mobiler Rechner, der davon ausgehen kann, daß ein anderer Rechner, mit dem er kommunizieren will, keine aktuelle Adreßinformation über ihn besitzt, ein sogenanntes *Binding Update* bei seinem Kommunikationspartner initiiert. Im allgemeinen kann er diesen jedoch nicht direkt informieren, da zwischen den beiden Rechnern in der Regel keine Beziehung zur Sicherung der Authentizität von Nachrichten besteht. Deshalb wird vorgeschlagen, daß der mobile Rechner eine *Binding Warning Message* bezüglich seines Kommunikationspartners an seinen *Home Agent* sendet, der dann wiederum eine *Binding Update Message* an den Kommunikationspartner schickt. Diese Meldung trifft in der Regel deutlich früher beim Kommunikationspartner ein, als wenn sie erst durch die erste Meldung vom Kommunikationspartner zum *Home Agent* ausgelöst wird. Darüberhinaus erhöht der Vorschlag die Robustheit von *Mobile IP*, da auch bei Verlust eines (beliebigen) einzelnen IP-Pakets dennoch mindestens ein *Binding Update* erfolgt.

Damit ein mobiler Rechner entscheiden kann, ob er bezüglich eines anderen Rechners eine *Binding Warning Message* senden soll, wird vorgeschlagen, daß er eine Liste der Partneradressen von in jüngerer Vergangenheit gesendeten oder empfangenen IP-Pakete hält. Befindet sich die Adresse eines zu sendenden IP-Pakets nicht auf dieser Liste, so wird bezüglich dieser Adresse eine *Binding Warning Message* an den *Home Agent* gesendet. Unter Umständen ist es sinnvoll, das gesamte Verfahren auf IP-Pakete zu beschränken, die TCP-Protokolldateneinheiten transportieren, da bei diesen zu erwarten ist, daß auch in Rückrichtung ein signifikanter Paketstrom stattfindet.

Das hier vorgeschlagene Verfahren zum frühen *Binding Update* ist in einer Reihe von Fällen von Nutzen. Diese werden weiter unten beschrieben. Zuerst soll jedoch gezeigt werden, daß das Verfahren im Normalfall im Vergleich zu gewöhnlichem *Mobile IP* mit *Route Optimization* keinen zusätzlichen Nutzen bringt. Gehen wir davon aus, daß der mobile Rechner (MN) zu einem Kommunikationspartner (CN) eine TCP-Verbindung aufbaut. Im Normalfall erreicht bei dem hier vorgeschlagenen Verfahren das frühe *Binding Update* den Kommunikationspartner zu spät, um zu verhindern, daß das erste IP-Paket vom Kommunikationspartner zum *Home Agent* (HA) geschickt wird. Sendet der Kommunikationspartner das zweite IP-Paket, so ist im

Normalfall bereits die reguläre *Binding Update Message* im Rahmen von *Mobile IP* mit *Route Optimization* eingetroffen. Um dies plausibel zu machen, nehmen wir an, daß die Laufzeiten von IP-Paketen zwischen Rechnern in einem IP-basierten Netz konstant und unabhängig von der Richtung sind. In Abbildung 4-1 sind die Bezeichnungen für die Laufzeiten zwischen den Knoten HA, MN und CN definiert. Die Laufzeit zwischen HA und MN heie t_1 , die Laufzeit zwischen MN und CN heie t_2 , und die Laufzeit zwischen CN und HA heie t_3 . Sie seien alle grer als Null.

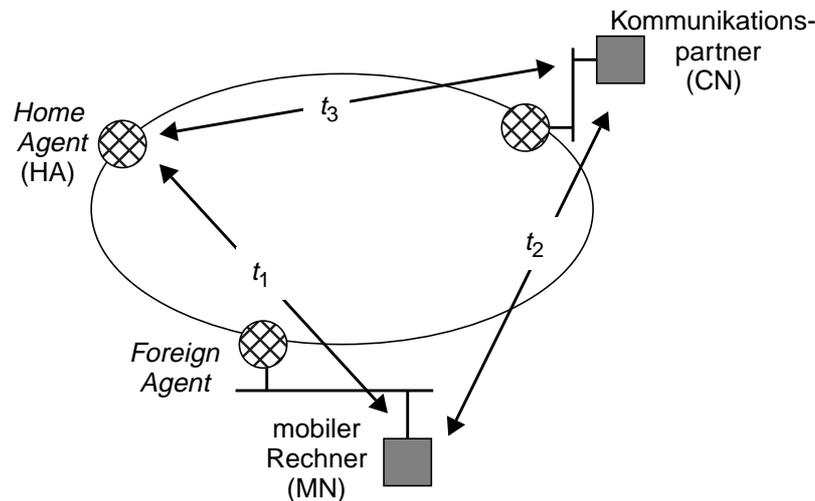


Abbildung 4-1: Benennung der Laufzeiten von IP-Paketen zwischen den beteiligten Knoten.

Abbildung 4-2 zeigt exemplarische Signalisier-Zeit-Diagramme, bei denen zum Zeitpunkt t_0 ein mobiler Rechner eine TCP-Verbindung zu einem Kommunikationspartner, der keine Ortsinformation ber den mobilen Rechner besitzt, aufbaut. Abbildung 4-2a zeigt die TCP-Signalisierung zum Verbindungsaufbau, wie sie bei gewhnlichem *Mobile IP* und bei *Mobile IP* mit *Route Optimization* stattfindet, und wie sie in aller Regel auch fr das frhe *Binding Update* abluft. Abbildung 4-2b zeigt die zugehrige Signalisierung von *Mobile IP* mit *Route Optimization*, wobei die zustzliche Signalisierung fr ein frhes *Binding Update* gestrichelt eingezeichnet ist.

Bearbeitungszeiten in den beteiligten Knoten werden vernachlssigt. Die erste TCP-Verbindungsaufbaumeldung erreicht den Kommunikationspartner zum Zeitpunkt $t_B = t_0 + t_2$. Sie wird sofort mit einer Besttigungsmeldung quittiert. Die frhe *Binding Update Message* trifft zum Zeitpunkt $t_I = t_0 + t_1 + t_3$ beim Kommunikationspartner ein. In der Regel kann man davon ausgehen, da $t_2 \leq t_1 + t_3$, da also die frhe *Binding Update Message* nicht rechtzeitig eintrifft, um zu verhindern, da die erste Besttigungsmeldung zuerst an den Home Agent des mobilen Rechners geschickt wird. Die erste Besttigungsmeldung lst beim HA eine regulre *Binding Update Message* aus, die den CN zum Zeitpunkt $t_{II} = t_0 + t_2 + 2t_3$ erreicht. Erreicht die TCP-Besttigungsmeldung den MN, so wird diese wiederum quittiert und in der Regel werden auch sofort Nutzdaten zur bermittlung vom CN angefordert, fr die frhestens zum Zeitpunkt $t_{ACK} = t_0 + t_1 + 2t_2 + t_3$ mit der bermittlung begonnen wird.

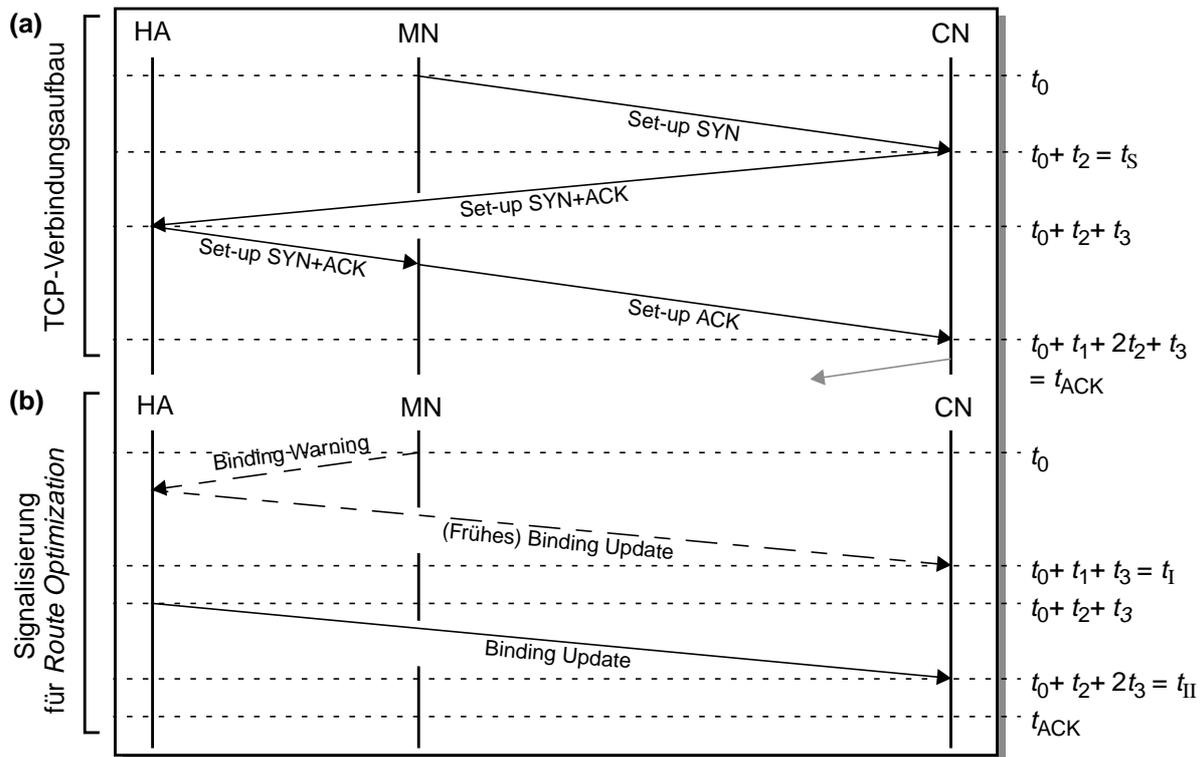


Abbildung 4-2: Signalisier-Zeit-Diagramme für (a) TCP-Verbindungsaufbau und (b) parallel erfolgende Signalisierung für *Mobile IP* mit *Route Optimization*, wobei Signalisierung für ein frühes *Binding Update* gestrichelt eingezeichnet ist.

Die frühe *Binding Update Message* trifft immer vor dem Beginn der Übertragung von Nutzverkehr beim CN ein, da

$$t_I < t_{ACK} \tag{4-1}$$

Geht man davon aus, daß sich die Laufzeiten t_1 , t_2 und t_3 zueinander wie die Seitenlängen eines Dreiecks verhalten, daß also eine der Laufzeiten nie die Summe der beiden anderen übersteigt, so kann man zeigen, daß auch die reguläre *Binding Update Message* vor der Übermittlung von Nutzverkehr beim CN eintrifft, da

$$t_{II} \leq t_{ACK} \tag{4-2}$$

und daß diese nicht vor der frühen *Binding Update Message* eintreffen kann, da

$$t_I \leq t_{II} \tag{4-3}$$

Zusammenfassend kann man sagen, daß unter den getroffenen Annahmen lediglich für den möglichen Sonderfall $t_{II} = t_{ACK}$ der Vorschlag eines frühen *Binding Update* von Nutzen ist, da in allen anderen Fällen ein reguläres *Binding Update* dasselbe leistet. Dennoch hat das vorgeschlagene Verfahren neben der schon genannten höheren Robustheit noch weitere Vorteile, die seinen Einsatz rechtfertigen können.

Für den Fall, daß ein mobiler Rechner in kurzer Folge beginnt, mehrere TCP-Verbindungen parallel zum selben Kommunikationspartner aufzubauen, was zum Beispiel beim gleichzeitigen Laden mehrerer Bilder auf einer WWW-Seite vorkommt, können die Verbindungsaufbau-

meldungen der zweiten oder dritten Verbindung unter Umständen bereits von der frühen *Binding Update Message* profitieren. Dies ermöglicht nicht nur eine sofortige direkte Wegführung, sondern reduziert dadurch auch die Zahl der regulären *Binding Update Messages*, die vom *Home Agent* als Folge der zu ihm gesendeten IP-Paketen geschickt werden.

Wie in den Kapitel 2.5.1 und 3.5.4 erläutert, muß zwischen *Home Agent* und Kommunikationspartner eine Sicherheitsbeziehung bestehen, bei der zum Schutz gegen Wiederverwendung von Nachrichten entweder Zeitstempel oder sogenannte *Nonces* verwendet werden. Haben die beiden Rechner lange Zeit nicht miteinander kommuniziert, so kann es sein, daß der Kommunikationspartner die erste Meldung vom *Home Agent* nicht auf ihre Erstverwendung überprüfen kann und deshalb der Wiederverwendungsschutz erst wieder synchronisiert werden muß. Solange, bis eine gesicherte *Binding Update Messages* den Kommunikationspartner erreicht, werden IP-Pakete über den *Home Agent* geführt. Auch hier kann das frühe *Binding Update* von Nutzen sein, da die Synchronisierung des Wiederverwendungsschutzes früher beginnen kann und damit die IP-Pakete auch früher auf direktem Weg geführt werden können.

Schließlich muß in Rechnung gestellt werden, daß die oben getroffenen Annahmen zu Laufzeiten in IP-basierten Netzen idealisiert sind. Üblicherweise sind sie nicht konstant und nicht unabhängig von der Richtung. Auch die Annahme, daß der direkte Weg zwischen zwei Knoten schneller ist als ein Umweg über einen dritten, ist nicht unbedingt garantiert. Für den Fall, daß beispielsweise $t_3 > t_1 + t_2$ ist, ergibt sich $t_{II} > t_{ACK}$, während nach Gleichung 4-1 weiterhin gilt, daß $t_I < t_{ACK}$ ist. Dies bedeutet, daß die reguläre *Binding Update Message* zu spät eintrifft, während die frühe *Binding Update Message* nach wie vor deutlich vor der Übermittlung von Nutzverkehr ankommt.

Abschließend soll gezeigt werden, daß im Zusammenhang mit einem Wechsel des *Foreign Agent* („handover“) das frühe *Binding Update* unter bestimmten Umständen früher eine optimale Wegführung ermöglicht. Besitzt der Kommunikationspartner nämlich noch ein *Binding* zum alten *Foreign Agent*, so verkürzt sich in der Regel die Zeit, bis zum Eintreffen der letzten Verbindungsaufbaumeldung (t_{ACK}), während sich die Zeit bis zum Eintreffen des regulären *Binding Update* (t_{II}) soweit erhöht, so daß in der Regel zuvor bereits mit der Übertragung von Nutzverkehr begonnen wurde. Das frühe *Binding Update* dagegen kann rechtzeitig genug kommen, um bereits für das erste Paket des Nutzverkehrs für direkte Wegführung zu sorgen. Abbildung 4-3 illustriert dies. Um das Bild übersichtlich zu halten, wurde weder für die Laufzeiten noch im Signalisier-Zeit-Diagramm zwischen altem und neuem Ort des mobilen Rechners (MN) unterschieden.⁴ In der Regel trifft das reguläre *Binding Update* zu spät ein, da $t_{ACK} < t_{II}$ für $t_2 < t_1 + t_3$. In jedem Fall trifft das frühe *Binding Update* früher ein ($t_I < t_{II}$). Wenn $t_1 + t_3 < 3t_2$ ist, dann trifft es bereits vor der Übertragung des ersten Nutzpaketes ein ($t_I < t_{ACK}$).

4. Würde man dies tun, so würde das frühe *Binding Update* im Vergleich zum regulären *Binding Update* noch besser abschneiden, da t_I ungefähr gleich bliebe, während t_{ACK} und t_{II} etwas größer würden.

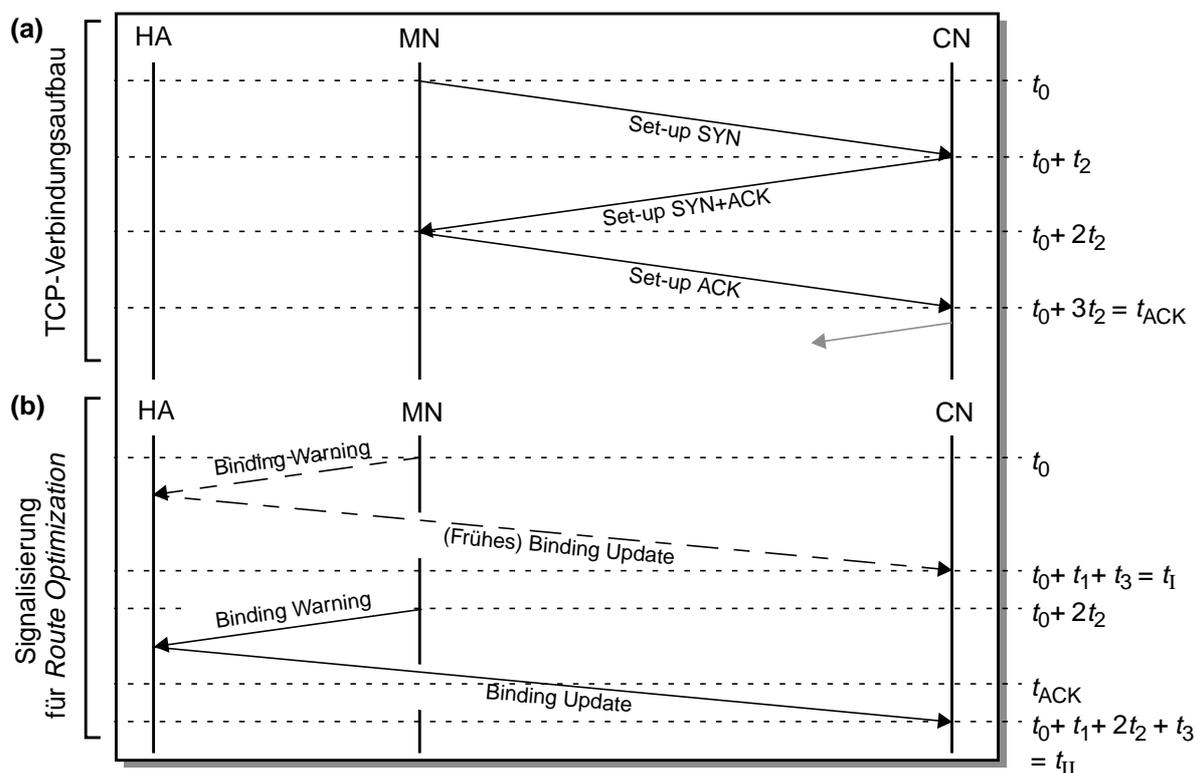


Abbildung 4-3: Signalisier-Zeit-Diagramme für (a) TCP-Verbindungsaufbau und (b) parallel erfolgende Signalisierung für *Mobile IP* mit *Route Optimization*

jeweils unter der Annahme, daß der Kommunikationspartner ein veraltetes *Binding* besitzt* (wobei Signalisierung für ein frühes *Binding Update* gestrichelt eingezeichnet ist).

* Weder für die Laufzeiten noch bei den Instanzen im Diagramm wird zwischen altem und neuem Ort des mobilen Rechners (MN) unterschieden.

Als Weiterentwicklung des Verfahrens wäre auch denkbar, beim Wechsel des *Foreign Agent* frühe *Binding Updates* bezüglich aller Kommunikationspartner, zu denen eine TCP-Verbindung besteht, zu initiieren.

4.2 Fortentwicklung der GSM-Mobilitätsverwaltung

SS7-basierte Mobilitätsverwaltung mit Hilfe einer zweistufigen Datenhaltung in HLRs und VLRs ist die dominierende Technik in zellularen Mobilfunksystemen. Nicht nur GSM-MAP sondern auch amerikanische Systeme wie AMPS (*Advanced Mobile Phone System*), IS-54 TDMA und IS-95 CDMA basieren auf einer gleichartigen Technik, wie sie im EIA/TIA⁵ Interim Standard 41 (IS-41) [Wil92] spezifiziert ist. Da sich nach einer Entscheidung von ETSI die UMTS-Mobilitätsverwaltung aus der von GSM/GPRS entwickeln soll, bleibt diese Technik auch für zukünftige Systeme von großer Bedeutung. In den folgenden Unterkapiteln werden einige Vorschläge zur Fortentwicklung der GSM-Mobilitätsverwaltung beschrieben und kurz

5. *Electronic Industry Association / Telephone Industry Association.*

qualitativ bewertet. Anders als bei den Vorschlägen in [MJ94] sollen jedoch zwei GSM-Prinzipien aufrecht erhalten werden:

- Die Antwort auf eine Anfrage wird immer direkt an die anfragende Instanz zurückgeschickt, da dies mit dem Konzept der entfernten Prozeduraufrufe des TCAP in Übereinstimmung ist.⁶
- Einem Teilnehmer ist nicht zu jedem Zeitpunkt eine exklusive Aufenthaltsrufnummer zugewiesen, da dies den Rufnummernbedarf dramatisch erhöhen würde.⁷

Die Unterkapitel 4.2.1 und 4.2.3 beschreiben im wesentlichen Vorschläge aus der Literatur. Eigene Vorschläge, die mit den Vorschlägen des jeweils vorausgehenden Unterkapitels verwandt sind, werden in den Unterkapiteln 4.2.2 und 4.2.4 beschrieben.

4.2.1 Zusätzliche Datenkopien

Einige Vorschläge versuchen die Kosten für Datenbankanfragen zu verringern, indem sie an potentiellen Nachfrageorten zusätzliche Datenkopien oder Information über die aktuelle Datenplatzierung bereithalten. Dadurch kann gegebenenfalls Signalisierung zwischen VLR und HLR eingespart werden.

Als eine einfache Variante dieses Vorgehens wird hier vorgeschlagen, an ausgewählten VLRs Datenkopien des Teilnehmerprofils zu halten. Beim Wechsel des zuständigen VLRs (siehe Abbildung 2-8 auf Seite 23) kann dann die Übermittlung des Teilnehmerprofils vom HLR zum VLR eingespart werden. Da das Teilnehmerprofil sich nicht besonders häufig ändert, bei Änderung jedoch aktualisiert werden muß, bietet es sich an, die Datenhaltung mit Hilfe eines *shadowing*-Verfahrens im *supplier update mode* durchzuführen. Die Entscheidung über die Platzierung von Datenkopien in VLRs kann im HLR getroffen werden, da dort mit Hilfe der eintreffenden Ortsaktualisierungsmeldungen eine globale Sicht auf das Bewegungsverhalten des Teilnehmers erlangt werden kann, so daß Kopien des Teilnehmerprofils in häufig besuchten VLRs plaziert werden können.⁸ Der Vorschlag bedarf Änderungen in den VLRs und in den HLRs sowie einer Erweiterung der D-Schnittstelle zur Unterstützung des *shadowing*-Verfahrens. Abwärtskompatibilität kann gewährleistet werden.

6. Das bedeutet, daß die Weitergabetechnik *Abgeben* nicht zugelassen ist, lediglich *Verweisen* und *Verketteten* sind gestattet.

7. Will man Signalisierung bei der Ortssuche reduzieren, so ist es sinnvoller, die ISUP-Rufaufbausignalisierung zu erweitern, so daß die gerufene Nummer (die MSRN) mitgeführt werden kann. Dadurch wird es (wie bei *Mobile IP*) möglich, mehrere Teilnehmer unter der selben Aufenthaltsrufnummer (*care-of address*) zu adressieren, was Signalisierung zur Beschaffung einer exklusiven Aufenthaltsrufnummer auf Rufbasis unnötig macht.

8. Alternativ ist denkbar, daß die Entscheidung zur Speicherung einer Datenkopie dezentral in den VLRs getroffen wird. Damit würde das Verfahren unabhängig davon, ob ein HLR grundsätzlich über Aufenthaltsortsaktualisierungen informiert wird (s. Kapitel 4.2.3).

In [JLM94] wird vorgeschlagen, in einem MSC/VLR-Knoten *Cache*-Kopien mit Information über den Aufenthaltsort von in jüngerer Vergangenheit gerufener Teilnehmer zu halten. Diese unter Umständen veraltete Information wird verwendet, um direkt vom besuchten VLR eines Teilnehmers eine Aufenthaltsrufnummer zu erfragen, ohne dies über dessen HLR tun zu müssen. Kann das befragte VLR keine Aufenthaltsrufnummer bereitstellen („*cache miss*“), weil der Teilnehmer sich von dort wegbewegt hat, so war die Signalisierung unnötig und es muß die übliche Anfrage via das HLR erfolgen. In [JLM94] wird nicht näher spezifiziert, in welcher Form die Information über den vermuteten Aufenthaltsort eines Teilnehmers erlangt wird. Will man jedoch bestehende GSM-Protokolle möglichst wenig verändern, so bietet sich an, diese Information aus der zuletzt erlangten Aufenthaltsrufnummer eines Teilnehmers abzuleiten.⁹ Das anfragende VLR muß sich dann gegenüber dem besuchten VLR lediglich wie ein anfragendes HLR verhalten. Für die eigentliche Mobilitätsverwaltung sind dann also Änderungen nur in den MSC/VLRs notwendig, in denen dieses *Cache*-Verfahren implementiert wird. Um eine Entscheidung treffen zu können, ob und in welchen MSC/VLRs *Cache*-Kopien gehalten werden sollen, sind aber je nach dem gewählten Vorgehen unter Umständen noch zusätzliche Modifikationen des GSM-Systems notwendig.

Eine *Cache*-Kopie an einem MSC/VLR-Knoten zu halten, lohnt sich nur, wenn die Einsparungen durch „*cache hits*“ die Zusatzkosten, die nach jeder Bewegung des Teilnehmers durch einen „*cache miss*“ entstehen¹⁰, übersteigen. [JLM94] definiert dazu für jeden Teilnehmer und jede MSC/VLR eine *local call-to-mobility ratio* (LCMR), die das Verhältnis der Rufrate von einer MSC/VLR zu einem Teilnehmer im Verhältnis zur Rate der Ortsaktualisierungen des gerufenen Teilnehmers setzt. Daneben hängt die Häufigkeit von „*cache hits*“ zu einem gewissen Grad auch von der Varianz der Aufenthaltszeiten in einem MSC/VLR ab [LH96]. Im allgemeinen lohnt sich das Verfahren nur bei Teilnehmern, die sich relativ zu der Rate kommender Rufe nur selten bewegen oder bei denen kommende Rufe ihren Ursprung relativ häufig am selben Ort haben. Letzteres gilt auch nur eingeschränkt, da bei einem Ruf, dessen Ursprung sich am aktuellen Aufenthaltsort eines Teilnehmers befindet, auch ohne Einsatz des *Cache*-Verfahrens die Signalisierung zur Zielsuche eingespart werden kann, indem vor der entfernten Zielsuche zuerst das lokale VLR befragt wird, ob der Teilnehmer zur Zeit dort eingebucht ist.

4.2.2 Einsatz der *Datacycle Architecture*

Im letzten Unterkapitel wurden zwei Verfahren vorgestellt, wie Kopien des Teilnehmerprofils und der Ortsinformation im Vorraus dort verfügbar gemacht werden können, wo sie nachgefragt werden. Die in Kapitel 3.5.5 vorgestellte *Datacycle Architecture* kann auf etwas andere Weise Vergleichbares leisten. Wendet man sie ausschließlich für HLR-Daten an und beläßt die

9. Darüberhinaus könnte die Information aktualisiert werden, wenn ein besuchender Teilnehmer eine MSC/VLR verläßt und das voraussichtliche Ziel ermittelbar ist – zum Beispiel wenn vor einer VLR-Eintragslöschung ein anderes VLR die IMSI und Authentisierungsparameter eines Teilnehmers erfragt.

10. Vernachlässigt wird bei dieser Überlegung der Fall, daß eine veraltete Information durch Rückkehr des Teilnehmers an den alten Ort wieder gültig wird.

verbleibende GSM-Datenhaltung, so ergibt sich ein System, bei dem Teilnehmerdaten periodisch auf einem *Broadcast*-Medium den MSC/VLRs zur Verfügung gestellt werden. Diese sind dadurch virtuell an jeder MSC/VLR in aktueller Form vorhanden. Die HLRs werden von individuellen Leseoperationen entlastet. Sie müssen nur noch individuelle Schreiboperationen durchführen und dazu ihren gesamten Datenbestand kontinuierlich auf den Kanälen des *Broadcast*-Mediums aussenden. VLRs müssen für Anfragen nicht mehr mit dem HLR kommunizieren. Innerhalb einer Zykluszeit stehen ihnen die entsprechenden Daten auf einem der Datenkanäle zur Verfügung. Das Teilnehmerprofil muß bei einer Ortsaktualisierung nicht transferiert werden und die Aufenthaltsrufnummer kann direkt vom besuchten VLR erfragt werden. Abbildung 4-4 zeigt das Prinzip der *Datacycle Architecture* angewendet auf ein GSM-System. Beispielhaft sind zwei Datenkanäle je HLR eingezeichnet. Weiß man in solch einem System, welcher Teilnehmer (bzw. welche IMSI oder MSISDN) welchem HLR zugeordnet ist und welche Datenkanäle von diesem HLR gespeist werden, so kann man gezielt nur diese Datenkanäle abhören. Noch besser ist es, wenn bekannt ist, welche Datenkanäle welchen Teilnehmerdaten zugeordnet sind, dann muß jeweils nur ein Datenkanal abgehört werden.

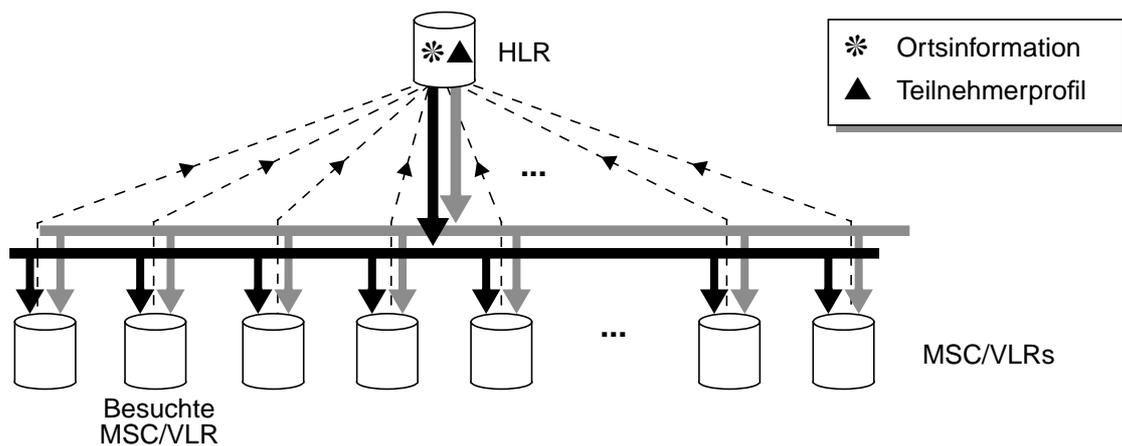


Abbildung 4-4: Verwendung der *Datacycle Architecture* in einem GSM-System.

Geht man davon aus, daß pro Teilnehmer etwa 1000 Byte Daten auf einem Datenkanal mit 2 Gbit/s gesendet werden, so kann man mit Daten von 100 000 Teilnehmern je Datenkanal eine Zykluszeit von etwa 0,4 Sekunden erzielen. Da die Ortsinformation im allgemeinen nur aus einigen wenigen Bytes besteht und da das Bereitstellen des Teilnehmerprofils weniger strikten Antwortzeitanforderungen unterliegt, kann man für das Teilnehmerprofil auch eine längere Zykluszeit (und u. U. andere Datenkanäle) vorsehen, während man sie für die Ortsinformation beibehält. Dann kann man noch deutlich mehr Teilnehmerdaten je Datenkanal versenden.

Der hier beschriebene Vorschlag kann zusätzlich in ein GSM-System eingebracht werden und mit dem bestehenden Standard-GSM-Verfahren koexistieren. Er bedarf jedoch bedeutender Änderungen in den beteiligten Netzknoten und zusätzlicher Netzinfrastruktur – z. B. Glasfaserkabel zur Realisierung des *Broadcast*-Mediums und zusätzlicher Hardware zum Beschicken und Abhören des *Broadcast*-Mediums. Es zementiert die zentrale Funktion der HLRs und skaliert nur schlecht, da jedes HLR jedes MSC/VLR auf einem *Broadcast*-Medium erreichen können muß. Das Verfahren zeichnet sich auch nicht gerade durch Datensparsamkeit aus, da

Information über einen Teilnehmer periodisch im gesamten Netz bekannt gemacht wird. Die Vorteile des Verfahrens bestehen darin, daß HLRs weniger Datenbankoperationen durchführen müssen und dadurch leistungsfähiger gemacht werden, daß das eigentliche Signalisieretz von Signalisierung entlastet wird, und daß die Antwortzeit von HLR-Anfragen in einem durch die Zykluszeit bestimmten Zeitintervall liegt, was sowohl die mittlere Antwortzeit als auch die Varianz derselben reduzieren kann. Somit kann es die Leistungsfähigkeit kleiner GSM-Netze erhöhen, ist jedoch in der Regel für große Netze nicht geeignet.

4.2.3 Verzicht auf Ortsaktualisierungen im HLR

Eine Reihe von Vorschlägen verzichten auf einen Teil der HLR-Aktualisierungen. Dadurch wird im allgemeinen Signalisierung zur Ortsaktualisierung eingespart, während die Kosten zur Zielsuche steigen. Dies lohnt sich in der Regel frühestens dann, wenn Ortsaktualisierungen nicht viel seltener erfolgen als Zielsuchen. Nachteil dieser Verfahren ist es, daß die Antwortzeiten der Zielsuche in der Regel vergrößert werden. Die Vorschläge lassen sich in zwei Gruppen einteilen.

In der ersten Gruppe wird bei einigen VLR-Wechseln gänzlich auf eine Ortsdatenaktualisierung verzichtet [TN95, Lin97]. Wird eine Aufenthaltsrufnummer benötigt, so muß das HLR gegebenenfalls bei mehreren VLRs anfragen. Die Verfahren unterscheiden sich darin, in welcher Weise die Befragung der VLRs organisiert wird, d. h. welche VLRs in welcher Reihenfolge – gegebenenfalls auch gleichzeitig – befragt werden. Die Problemstellung ist vergleichbar mit der, die sich beim Ausrufen (*Paging*) eines Teilnehmers in mehreren Zellen eines Aufenthaltsgebiets ergibt [Pla94, HA95, Tab95, AHL96]. Gruppiert man VLRs in statischen Gruppen, so ist keine Änderung der GSM-Schnittstellen notwendig. In allen anderen Fällen muß in der Regel zusätzliche Information im System verteilt werden.

Die zweite Gruppe von Vorschlägen aktualisiert zwar die Ortsinformation im HLR nicht, hinterläßt dafür jedoch an anderer Stelle einen Verweis auf das aktuell besuchte VLR. Diese Verfahren sind besonders dann von Vorteil, wenn ein Teilnehmer sich fern von seinem HLR befindet und Signalisierung zur Verwaltung von Verweiszeigern günstiger ist als Signalisierung zum entfernten HLR. In [JL95, JLM95] wird am jeweils zuletzt besuchten VLR ein Weiterleitungszeiger hinterlassen. Durch Bewegung zwischen VLRs entsteht dabei eine Kette von Weiterleitungszeigern (die bei Schleifenbildung verkürzt wird). Droht die Länge der Kette einen gewissen Wert zu übersteigen, so wird das HLR über das aktuell besuchte VLR informiert und die Kette von Weiterleitungszeigern aufgelöst. Bei der Zielsuche wird die Anfrage vom HLR an das zuletzt gemeldete VLR gestellt und von dort entlang der Kette weitergereicht. In [JL95] geschieht dies mit Hilfe der Weitergabetechnik *Abgeben*. Da dies dem für dieses Kapitel aufgestellten Prinzip widerspricht, Anfragen immer an den Anfrager zurückzureichen (siehe S. 94), wird deshalb hier vorgeschlagen, die Weitergabetechnik *Verketteten* zu verwenden. Zur Realisierung des Verfahrens muß das HLR nicht modifiziert werden. Lediglich zwischen den VLRs muß zusätzliche Signalisierung zur Verwaltung der Weiterleitungszeiger erfolgen.

Das Endgerät muß, wie in GSM-Systemen üblich, dem neuen VLR Informationen übermitteln, die es ermöglichen, das zuletzt besuchte VLR zu identifizieren. Das Verfahren kann abwärtskompatibel gestaltet werden und damit mit nicht-erweiterten VLRs koexistieren.

In [HA96] wird ein sogenannter „lokaler Anker“ (*local anchor*) vorgeschlagen. Es wird nur ein einziger Weiterleitungszeiger in einem der zuletzt besuchten VLRs hinterlassen, der anstelle des HLRs über VLR-Wechsel informiert wird. Dadurch wird günstige lokale VLR-VLR-Signalisierung ermöglicht und die Zielsuche lediglich um einen einzelnen Verkettungsschritt verlängert. [HA96] sieht zwei Varianten vor. Beim *statischen lokalen Anker* fungiert immer das VLR als lokaler Anker, an dessen MSC zuletzt ein Ruf eingegangen ist. Das heißt der lokale Anker bleibt solange fest, bis eine Zielsuche stattfindet. Die Zielsuche gibt den Anstoß zur Verlagerung des lokalen Ankers zum aktuell besuchten VLR. Damit soll erreicht werden, daß das Verfahren sich bei einer hohen Rate an Zielsuchen nicht viel schlechter verhält als Standard-GSM. Beim *dynamischen lokalen Anker* wird die beschriebene Vorgehensweise erweitert, indem eine Verlagerung des lokalen Ankers zum aktuell besuchten VLR auch bei einem beliebigen VLR-Wechsel zugelassen wird. Zur Realisierung des Verfahrens muß das HLR nicht modifiziert werden. Die Information, welches VLR bei einem VLR-Wechsel als lokaler Anker fungiert, muß von VLR zur VLR weitergereicht werden. Zum Löschen eines Eintrags in einem besuchten VLR, das nicht als lokaler Anker fungiert, und zum Aktualisieren des lokalen Ankers können zwischen den VLRs MAP-Meldungen der D-Schnittstelle¹¹ auf der G-Schnittstelle verwendet werden. Auch hier gilt, daß das Verfahren abwärtskompatibel gestaltet werden kann und damit mit Standard-VLRs koexistieren kann.

4.2.4 Einführung von *Proxy HLRs*

In [HMS97, MH98] haben wir ein Verfahren vorgeschlagen, das in vielen Aspekten dem Verfahren mit lokalem Anker sehr ähnlich ist. Wir haben es als Verfahren mit *Mirror HLR* bezeichnet und verschiedene Varianten unterschieden. Auch hierbei besteht die Idee darin, einen einzelnen zusätzlichen Weiterleitungszeiger im System zu halten. Anders als der lokale Anker muß dieser Weiterleitungszeiger nicht notwendigerweise in einem jüngst besuchten VLR plaziert sein. Die Verlagerung kann zu einem beliebigen VLR und sogar zu einem beliebigen dafür vorgesehenen Netzknoten erfolgen. Eine Sonderform hiervon stellt ein Verfahren dar, das jedem VLR für bestimmte Teilnehmer fest einen Netzknoten zuweist, in dem die Weiterleitungszeiger gehalten werden. In dieser Arbeit, ebenso wie in [SM98], soll dieser spezielle Netzknoten als *Proxy HLR*¹² bezeichnet werden. Ein *Proxy HLR* kann ein gewöhnliches HLR

11. Zur Benennung der Schnittstellen siehe Abbildung 2-7 auf Seite 21.

12. Der Begriff *proxy HLR* wird auch in [DVS97] verwendet. Dort wird vorgeschlagen, die Ortsdatenverwaltung ohne den Einsatz von VLRs (*flat location management*) durchzuführen. Nur in besonderen Fällen wird eine zweistufige Hierarchie zugelassen. Die Knoten in der unteren Stufe dieser Hierarchie werden als *proxy HLRs* bezeichnet.

sein, das zusätzlich als *Proxy HLR* fungiert. Konzeptionell handelt es sich um einen zusätzlichen Knoten, der die GSM-Datenhaltungshierarchie um eine Stufe erweitert. Dadurch wird die Skalierbarkeit der GSM-Mobilitätsverwaltung erhöht. Abbildung 4-5 zeigt dies exemplarisch.

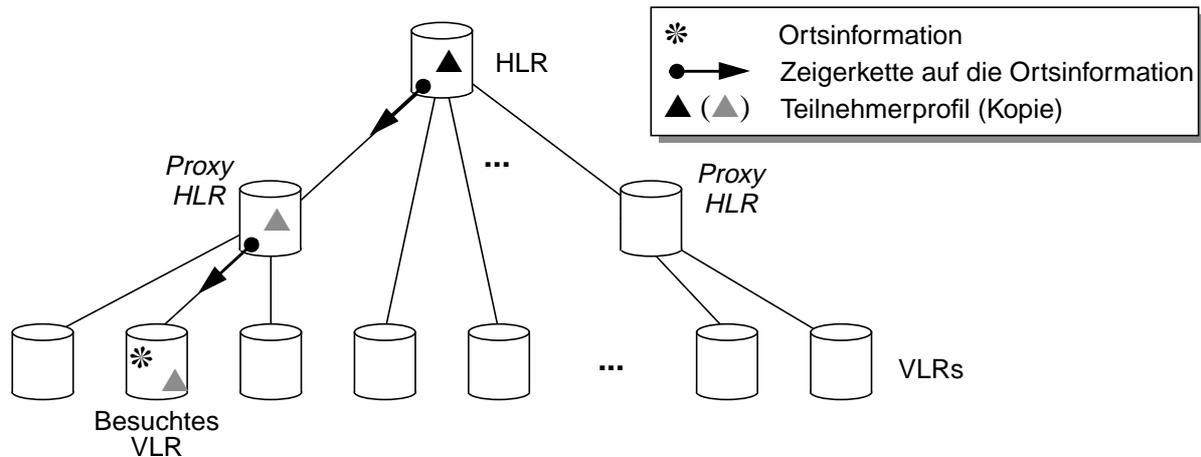


Abbildung 4-5: Logische Struktur des teilnehmerspezifischen Datenbanksystems von GSM, die an einigen Stellen um *Proxy HLRs* erweitert ist.

Ein *Proxy HLR* verhält sich einem VLR gegenüber wie ein HLR. Es hält eine Kopie des Teilnehmerprofils und stellt es auf Anfrage zur Verfügung. Dem eigentlichen HLR gegenüber verhält es sich wie ein VLR, bei dem ein Teilnehmer sich eingebucht hat. Solange dieser Teilnehmer lediglich zwischen den dem *Proxy HLR* zugeordneten VLRs wechselt, erreicht das HLR keine Aufenthaltsortsaktualisierung. Dies macht die Aufenthaltsortsinformation im HLR ungenauer, so daß ein (geringfügig) verbesserter Schutz von personenbezogenen Daten erreicht wird, da ein durch Beobachten der Information im HLR erstelltes Bewegungsprofil eines Teilnehmers weniger aussagekräftig ist.

Die beteiligten HLRs und VLRs müssen dazu nicht verändert werden. Lediglich die Adressierung im Signalisierungssystem Nummer 7 muß so modifiziert werden, daß die *Global Title Translation* in den VLRs die Signalisierung, die für das HLR eines Teilnehmers bestimmt ist, zum zuständigen *Proxy HLR* leitet. Damit kann im Prinzip für jeden Teilnehmer – oder zumindest für Gruppen von Teilnehmern – an jedem VLR individuell festgelegt werden, ob ein *Proxy HLR* für ihn verwendet werden soll oder nicht, und wenn ja, welches.

Geht man davon aus, daß Signalisierung zwischen VLR und *Proxy HLR* günstiger ist, als Signalisierung zwischen VLR und HLR, so sind Einsparungen in zweifacher Hinsicht möglich. Zum einen benötigen Aufenthaltsortsaktualisierungen zwischen VLRs unterhalb desselben *Proxy HLR* nur Signalisierung zwischen den beteiligten VLRs und dem *Proxy HLR*. Zum anderen werden Anfragen nach der Aufenthaltsrufnummer eines Teilnehmers von VLRs unterhalb des *Proxy HLR*, an dem sich der gerufene Teilnehmer aktuell befindet, ohne Beteiligung des HLR über das *Proxy HLR* an das besuchte VLR weitergereicht. Lediglich Zielsuchen von anderen VLRs laufen unter Beteiligung des HLR – und gegebenenfalls eines weiteren *Proxy HLR* – ab. Werden die *Proxy HLRs* so platziert, daß im Vergleich zu direkter HLR-VLR-Kommunikation keine oder nur geringe zusätzliche Kosten entstehen, wenn die Signalisierung über sie

geführt wird, so reduziert das Verfahren in jedem Fall die Signalisierlast im System und die Antwortzeiten für eine Reihe von Prozeduren. Dies wird durch eine erhöhte Belastung der Datenbankknoten und durch längere Antwortzeiten für Aufenthaltsortsanfragen von VLRs, die sich nicht in der Nähe des gerufenen Teilnehmers befinden, erkauft.

Ein sehr ähnlicher Vorschlag für die Unterstützung von internationalem „Roaming“ wird in [LL97a, LL97b, GSG97] als *Replicated Database* bezeichnet. Ohne daß dies dort explizit gesagt wird, setzt der Vorschlag das Dienstmerkmal *Support of Optimal Routeing* der GSM-Phase 2+ voraus, da auch für Teilnehmer fremder GSM-Netze eine Zielsuche initiiert wird.

Das *Proxy-HLR*-Verfahren stellt in der Tat eine ideale Ergänzung des Dienstmerkmal *Support of Optimal Routeing* (SOR) [G02.79] dar. Die Vorteile des *Proxy HLR* bei Aufenthaltsortsaktualisierungen und Ortsanfragen kommen voll zum tragen. Darüberhinaus kann gegebenenfalls auch SOR-spezifische Signalisierung eingespart werden. In der Phase 1 von SOR ist festgelegt, daß bei Rufen zwischen mobilen Teilnehmern die optimale Wegführung nur dann erfolgt, wenn beide Teilnehmer sich im selben Land befinden.¹³ Ist dies nicht der Fall, so wird bei SOR nach [G03.79] bei Rufen aus fremden Ländern die Ortsanfrage vom HLR negativ beantwortet und der Ruf wird – wie bei einfachem GSM – zuerst ins Heimatnetz geführt. Richtet man *Proxy HLRs* so ein, daß ein Teilnehmer in einem fremden Land (unabhängig davon, ob er dort ein Netz besucht) genau einem *Proxy HLR* in diesem Land zugeordnet ist, so läßt sich auch die unergiebigere Anfrage beim fernen HLR von einem *Proxy HLR* abfangen und unmittelbar negativ beantworten. Dadurch wird im Vergleich zur Phase 1 von SOR sowohl Signalisierung eingespart als auch die Rufaufbauzeit für Rufe zu Teilnehmern fremder GSM-Netze verringert.

4.3 Netztechnikenübergreifende Mobilitätsunterstützung

Für einen Nutzer von Kommunikationsdiensten ist die unterstützende Netztechnik in der Regel von nachrangiger Bedeutung. Er kann erwarten, daß Kommunikationsdienste an unterschiedlichen Orten und in unterschiedlichen Netzen in ähnlicher Weise genutzt werden können. Insbesondere im Bereich der personenbezogenen Telekommunikation ist es wünschenswert, daß die Mobilität eines Teilnehmers netztechnikenübergreifend unterstützt wird. So sollte beispielsweise Sprachkommunikation unabhängig davon stattfinden können, ob ein Teilnehmer momentan an einem Festnetzanschluß, an einem Mobilfunkendgerät oder an einem IP-Telefon im Internet erreichbar ist. Will man derartige personenbezogene, mobile Kommunikation in effizienter Weise realisieren, so ist es notwendig, Konzepte zur Dienstekonvergenz zu entwick-

13. Dies ist eine sehr vernünftige Einschränkung, da anderenfalls Aufenthaltsortsinformation über einen Teilnehmer von Netzen erlangt werden kann, die sich weder im Heimatland noch im momentan vom Teilnehmer besuchten Land befinden. Man denke etwa an den Fall, bei dem ein Teilnehmer aus Land X sich sowohl in Land Y als auch in Land Z aufhalten will, obwohl Y und Z so verfeindet sind, daß sie ihren Besuchern Kontakte mit dem jeweils anderen Land untersagen wollen und deshalb versuchen, Informationen über die Reisegewohnheiten der Teilnehmer aus Land X zu erlangen.

keln. Im Folgenden werden dazu zwei Vorschläge für effiziente netztechnikenübergreifende Mobilitätsunterstützung präsentiert. Der erste Vorschlag befaßt sich mit Konzepten zum effizienten Zusammenwirken zwischen einem IN-basierten UPT-Dienst und der GSM-Mobilitätsverwaltung. Der zweite Vorschlag sieht eine netzbetreiber- und netztechnikübergreifende Datenhaltung zur personenbezogenen Aufenthaltsortsverwaltung vor.

4.3.1 Zusammenwirken von UPT und GSM

Wir setzen hier eine sehr einfache Form eines UPT-Dienstes voraus. Ein Teilnehmer erhält von einem Festnetzbetreiber eine persönliche Rufnummer. Er kann sich an beliebigen Endgeräten innerhalb des Netzes registrieren, um dort Rufe absetzen und entgegennehmen zu können. Dieser Dienst wird in der Regel mit IN-Technik realisiert, wozu Rufe und Registrierungen durch SSPs unterstützt werden müssen. Gehen wir nun davon aus, daß die Verfügbarkeit dieses Dienstes auf eines oder mehrere GSM-Netze ausgedehnt und folgende Anforderungen erfüllt werden sollen:

- ❑ Rufe aus dem GSM-Netz zu einem UPT-Teilnehmer sollen möglichst direkt geführt werden.
- ❑ Der UPT-Teilnehmer soll sich mit einem eigenen SIM im GSM-Netz einbuchen können und dabei gleichzeitig eine UPT-Registrierung vornehmen.

In [Sch97b, Goy98] werden dazu drei Varianten untersucht:

- ❑ Variante A: GSM-Netz und die IN-Dienstlogik interagieren nicht.
- ❑ Variante B: Die UPT-Dienstlogik nutzt geschickt bestehende Mechanismen im GSM, um ein effizientes Zusammenwirken zu erreichen.
- ❑ Variante C: In das GSM-Netz wird zusätzliche Dienstlogik eingebracht, die effizient mit der UPT-Dienstlogik im Heimatnetz zusammenarbeitet.

Variante A führt zu Realisierungen, wie sie auch für einen Zugang zum UPT-Dienst aus fremden Festnetzen, die keine UPT-Unterstützung bieten, typischerweise gewählt werden. Registrierungen und gehende Rufe werden mit Hilfe von entgeltfreien Rufnummern an einen SSP im Heimatnetz des Teilnehmers geführt und dort wie netzinterne Rufe und Registrierungen behandelt. Das Einbuchen in das GSM-Netz kann nur mit einem GSM-SIM erfolgen. Kom-mende Rufe werden im Heimatnetz weitergeleitet. Keine der beiden oben genannten Anforderungen kann damit erfüllt werden.

Bei Variante B wird ein SCP im Heimatnetz als „virtuelles HLR“ [Sch97b] betrieben. Der Teilnehmer erhält ein SIM vom UPT-Anbieter. Für das besuchte GSM-Netz scheint er ein Teilnehmer eines fremden GSM-Netzes zu sein. Sollen Rufe aus dem GSM-Netz direkt geführt werden können, so muß im Rahmen des Dienstmerkmals *Support of Optimal Routeing* die Wahl einer UPT-Rufnummer eine Aufenthaltsrufnummernanfrage an das „virtuelle HLR“ auslösen. Diese Variante ist sehr elegant, da keine IN-Instanzen und keine Änderungen der Protokolle im GSM-Netz notwendig sind.

Die Variante **C** bedarf umfassender Änderungen im GSM-Netz, während die UPT-Infrastruktur im Heimatnetz nahezu unverändert bleiben kann. SSP-Funktionalität im GSM-Netz unterstützt kommende und gehende Rufe. Soll der UPT-Nutzer sich mit einem SIM einbuchen können, so sollte dieses vom GSM-Anbieter stammen und das HLR sollte die Registrierung selbsttätig beim SCP im Heimatnetz vornehmen können.

Variante **B** ist sicherlich die vielversprechendste, da im besuchten GSM-Netz keine UPT-spezifischen Erweiterungen notwendig sind und damit die Verfügbarkeit des UPT-Dienstes sehr einfach auf mehrere GSM-Netze ausgedehnt werden kann. Darüberhinaus verbleibt die Dienststeuerung beim SCP im Heimatnetz, was zum Beispiel eine einfache Erweiterung der UPT-Dienstlogik um neue Dienstmerkmale möglich macht. Dies gilt besonders dann, wenn das GSM-Netz CAMEL unterstützt. In [Sch97b, Goy98] wurde dazu beispielsweise untersucht, wie bei Registrierung an mehreren Endgeräten sichergestellt werden kann, daß bei Nichterreichbarkeit eines GSM-Endgeräts stattdessen ein anderes Endgerät ausgewählt wird.

4.3.2 Netzeübergreifende Aufenthaltsortsverwaltung

Die Mehrzahl der bisher vorgestellten Verfahren zur Aufenthaltsortsverwaltung profitieren von der Lokalität von Rufen und von Bewegung, d. h. davon,

- daß Rufe zu einem Teilnehmer ihren Ursprung häufig relativ nahe am aktuellen Aufenthaltsort oder relativ nahe am Heimatort eines Teilnehmers haben, und davon,
- daß ein Aufenthaltsortswechsel in der Regel zwischen Netzknoten stattfindet, die sich geographisch nahe sind.

In einer Welt, in der Kommunikationsdienste von einer Vielzahl von Netzbetreibern angeboten werden und auf unterschiedlichen Netztechniken beruhen, kann davon unter Umständen nicht mehr unmittelbar profitiert werden. Jeder Netzbetreiber hat in der Regel eigene Infrastruktur, und Interaktion zwischen den verschiedenen Netzen erfolgt nur an relativ wenigen Punkten. Dies führt dazu, daß zum Beispiel die Zielsuche bei Rufen zu Teilnehmern im Netz eines anderen Netzbetreibers nicht von der geographischen Nähe des gerufenen Teilnehmers profitieren können. Echte personenbezogene Kommunikation erlaubt es dem Teilnehmer, häufig das ihn unterstützende Netz zu wechseln. Denkbar ist daneben ein Szenario, bei dem ein Teilnehmer (auch unter Beibehaltung des verwendeten Endgeräts) aus einem IP-basierten Firmennetz in ein öffentliches GSM-Netz wechselt, um kurz darauf an seinem Festnetzanschluß zuhause erreichbar zu sein.¹⁴ Obwohl die Bewegung dabei geographisch relativ lokal erfolgt, kann die dadurch ausgelöste Signalisierung zur Aktualisierung der Aufenthaltsortsdaten davon in der Regel nicht profitieren.

14. Das Szenario läßt sich leicht um weitere Netze erweitern, so z. B. um satellitengestützte Kommunikation, UMTS-Netze und öffentliche schnurlose Festnetzzugänge.

Nimmt man noch den Fall hinzu, daß ein Teilnehmer gleichzeitig in mehreren Netzen eingebucht sein kann und bei kommenden Rufen die Dienstlogik, der Anrufer und/oder der gerufene Teilnehmer entscheiden können sollen, an welches Endgerät der Ruf geführt werden soll, wird offensichtlich, daß eine Aufenthaltsortsverwaltung, die gemeinsam für mehrere Netze erfolgt und unabhängig von der jeweiligen Netztechnik ist, sowohl funktional als auch bezüglich ihrer Leistungsfähigkeit Vorteile verspricht. Die Grundidee besteht darin, daß Netzknoten verschiedener Netze, die Teilnehmer innerhalb eines bestimmten geographischen Gebiets versorgen, vom selben Datenbankknoten versorgt werden. Sowohl Ortsaktualisierungen als auch Zielsuchen werden an diesen Knoten gerichtet. Denkbar wäre zum Beispiel ein VLR, das für alle Vermittlungsstellen aller Netzbetreiber eines Ortes zuständig ist. In [Sch96] haben wir eine Aufenthaltsortsverwaltung vorgeschlagen, bei der ein gemeinsames verteiltes Datenbanksystem (*Common Distributed Database*) genutzt wird, das wie im Vorschlag von [Wan93] organisiert ist, jedoch Zielsuche und Wegführung trennt. Da dieses Verfahren stark von der Lokalität von Rufen und Bewegung profitiert, erscheint es ideal für die vorgeschlagene Anwendung. Abbildung 4-6 zeigt das Prinzip der *Common Distributed Database* anhand zweier Netze, die das gemeinsame Datenbanksystem nutzen.

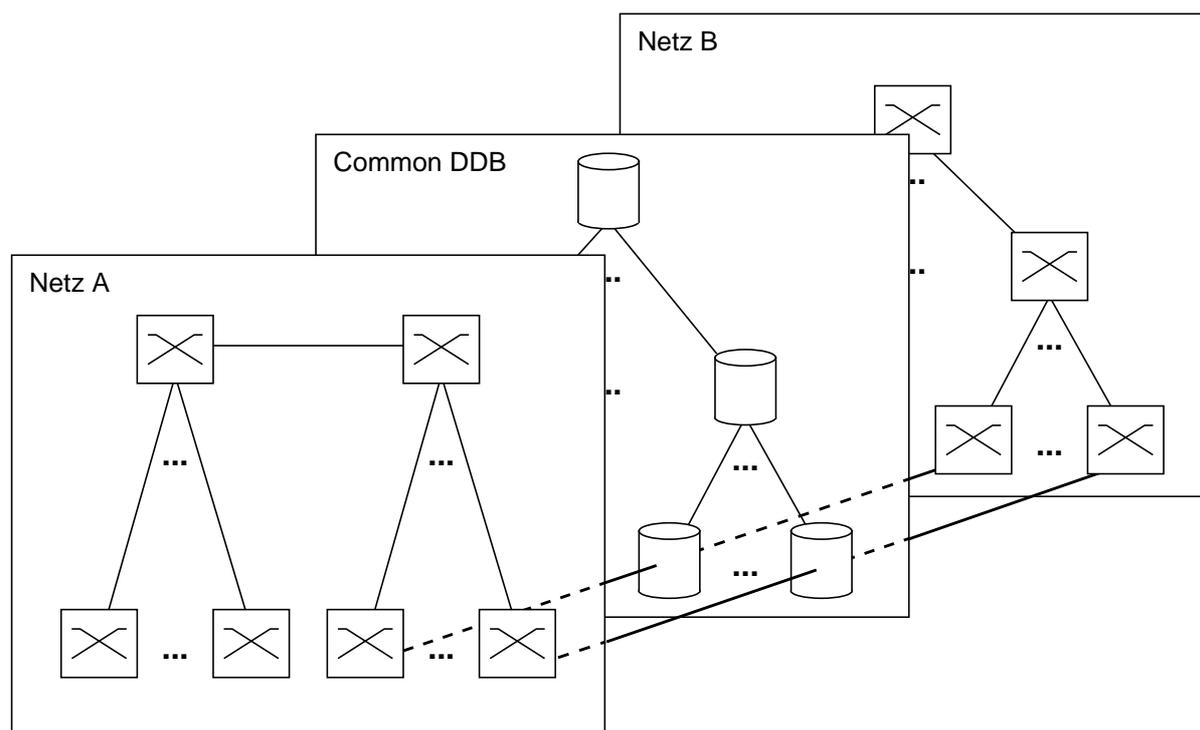


Abbildung 4-6: Prinzip der *Common Distributed Database*.

4.4 Zwischenbilanz der Arbeit

Dieses Kapitel beschließt den ersten Teil der Arbeit, in dem Mobilitätsunterstützung systematisiert und eine Reihe von Systemen und Vorschlägen präsentiert wurden. Die Bewertung der verschiedenen Vorschläge geschah dabei in allgemeinen in qualitativer Form und basierte an einigen Stellen auf Einsichten, die der Anschauung entnommen wurden. Dabei hat sich häufig

angedeutet, daß das Verhalten der Teilnehmer und die netztechnischen Voraussetzungen die Leistungsfähigkeit der Verfahren maßgeblich beeinflussen können. Auf diesem Hintergrund wird im zweiten Teil der Arbeit eine Vorgehensweise entwickelt, die basierend auf aus einem Teilnehmermodell gewonnenem Modellverkehr quantitative Vergleiche zwischen verschiedenen Netzstrukturen und Verfahren zur Mobilitätsverwaltung ermöglicht. Damit können sowohl deren Leistungsfähigkeit besser abgeschätzt, als auch Empfehlungen für die Anordnung und Dimensionierung von Netzelementen gewonnen werden.

Kapitel 5

Modellierung

Will man verschiedene Verfahren zur Mobilitätsverwaltung vergleichen und einer Leistungsbeurteilung unterziehen, so müssen die Auslöser von Mobilitätsverwaltungsprozeduren, der Ablauf dieser Prozeduren und die sie ausführenden und unterstützenden Instanzen geeignet modelliert werden. Dies geschieht in der vorliegenden Arbeit mit Hilfe eines Teilnehmermodells, eines Lastmodells und eines Netzmodells. Aus dem Teilnehmermodell ergibt sich, welche Mobilitätsverwaltungsprozeduren wie häufig und in welcher Weise angestoßen werden. Mit Hilfe des Lastmodells wird beschrieben, in welcher Weise diese Prozeduren ablaufen und welche Last in Form von Prozessorbelegung und zu übertragenden Nachrichten an welchen Stellen anfällt. Im Netzmodell schließlich wird dargestellt, in welcher Weise die anfallende Last verarbeitet wird, so daß schließlich im Rahmen einer Leistungsbewertung charakteristische Größen, wie zum Beispiel Antwortzeiten, ermittelt werden können.

Eine mobile Instanz (eine Person oder ein Endgerät), deren Verhalten für die Mobilitätsverwaltung eines Kommunikationsnetzes relevant ist, wird im weiteren als *Teilnehmer* bzw. als *mobiler Teilnehmer* bezeichnet. Das Teilnehmermodell wird in ein Bewegungsmodell (Unterkapitel 5.1) und in ein Rufmodell (Unterkapitel 5.2) unterteilt. Im Bewegungsmodell wird das Bewegungsverhalten von mobilen Teilnehmern, im Rufmodell werden die Versuche, Kommunikationsbeziehungen zu mobilen Teilnehmern aufzubauen (was hier als „Ruf“ bezeichnet wird), modelliert. Mehrere Teilnehmer werden zu Teilnehmerklassen zusammengefaßt. Verschiedene Teilnehmerklassen unterscheiden sich durch unterschiedliche Parameter der Bewegungs- bzw. der Rufmodelle. Die Beschreibung und die Parametrisierung der Modelle beziehen sich auf einen repräsentativen Teilnehmer einer Klasse.¹ Reale Teilnehmer können sich in ihrem Verhalten von solch einem repräsentativen Teilnehmer unterscheiden. Betrachtet werden soll jedoch das aggregierte Verhalten aller Teilnehmer einer Klasse. Daher wird angenommen, daß das Verhalten eines einzelnen Teilnehmers im Modell dem Verhalten eines repräsentativen Teilnehmers seiner Teilnehmerklasse entspricht. Ziel dieser Modellierung ist es also nicht, das Verhalten eines einzelnen realen Teilnehmers nachzubilden, sondern Modellverkehr zu erzeugen, der das aggregierte Verhalten aller Teilnehmer nachbildet und der als Referenz zur Bewertung verschiedener Systeme herangezogen werden kann.

1. Bei den Parametern der Modelle dient der Index ^(k) zur Bezeichnung der Klassenzugehörigkeit. Im folgenden wird an einigen Stellen auf diesen Index verzichtet, um die Notation einfach und übersichtlich zu halten.

Das Lastmodell (Unterkapitel 5.3) wird in Anlehnung an das konzeptionelle Modell des „Intelligenten Netzes“ (INCM) entwickelt. Es wird durch Abbildungen zwischen den Ebenen des INCM zunehmend verfeinert. Bis einschließlich der Abbildung auf die verteilte funktionale Ebene erfolgt die Lastbeschreibung unabhängig von der Netzarchitektur und der tatsächlichen Datenverteilung. Erst bei der Abbildung auf die physikalische Ebene werden die Netzarchitektur, zusätzlich notwendige Protokollfunktionen zur Kommunikation zwischen Netzknoten, Wegführung, Meldungslängen und Bearbeitungszeiten in bestimmten Typen von Prozessoren berücksichtigt. Dabei wird aus wenigen Szenarien der verteilten funktionalen Ebene eine Vielzahl von Szenarien der physikalischen Ebene. Der hier vorgestellte Ansatz fällt in die Gruppe der Verkehrsmodelle zur Beschreibung von Signalisierverkehr „Intelligenter Netze“ mit Hilfe von vorgegebenen Szenarien. In [VH98] werden solche Verkehrsmodelle in zwei Gruppen eingeteilt. Die erste Gruppe ist dadurch charakterisiert, daß sie den Meldungsverkehr anhand von Verteilungsfunktionen beschreibt, während die zweite Gruppe diesen mit Hilfe von vorgegebenen Szenarien („*pre-determined scenarios*“) erzeugt.

Bereits ein sehr einfaches Netzmodell (Unterkapitel 5.4), bestehend aus Übertragungsstrecken und Prozessoren, genügt, um mit Hilfe der aus dem Teilnehmermodell gewonnenen Raten für die verschiedenen Szenarien des Lastmodells das Verkehrsangebot – und damit indirekt auch die Auslastung der Netzkomponenten – zu bestimmen. Für Planung und Dimensionierung eines Netzes und seiner Knoten stellen diese Informationen eine wichtige Grundlage dar. Will man darüberhinaus Aussagen über die Durchlaufzeiten von Anforderungen durch Netzknoten oder die Ende-zu-Ende Antwortzeiten von Mobilitätsverwaltungsprozeduren gewinnen, so muß das Netzmodell verfeinert werden. Hierzu muß insbesondere die Ressourcenzuweisung innerhalb eines Netzknotens, das Multiplexen von Signalisiernachrichten über eine Übertragungsstrecke und ihre Verzögerung durch Laufzeiten nachgebildet werden. Die Ressourcenzuweisung innerhalb eines Netzknotens wird durch ein zweistufiges Verfahren modelliert, bei dem für Anforderungen zuerst ein zuständiger Prozessor ausgewürfelt wird und dann alle Anforderungen zum selben Prozessor auf Basis einer Prioritätenvergabe mit Unterbrechungsdistanz Bearbeitungszeit exklusiv zugewiesen bekommen.

5.1 Das Teilnehmermodell: Bewegung

In den folgenden Unterkapiteln wird auf das Modell des Bewegungsverhaltens von mobilen Teilnehmern näher eingegangen. Unterkapitel 5.1.1 stellt die Grundlagen des Bewegungsmodells dar. Die Unterkapitel 5.1.2, 5.1.3 und 5.1.4 beschreiben das stationäre Modell sowie Methoden zu seiner Parametrisierung und zur Überlagerung mehrerer identischer, unabhängiger Modelle. Im Unterkapitel 5.1.5 werden einfache Ansätze zur Kopplung des Bewegungsverhaltens von Teilnehmern und der Modellierung transienten Verhaltens vorgestellt.

5.1.1 Grundlagen

5.1.1.1 Markoffsche und Semi-Markoffsche Zustandsprozesse

Aus Sicht der Mobilitätsverwaltung läßt sich die Teilnehmerbewegung als ein diskreter stochastischer Prozeß in kontinuierlicher Zeit beschreiben. Im endlichen Zustandsraum dieses Prozesses ist dabei jedem möglichen Aufenthaltsort eines mobilen Teilnehmers genau ein Zustand zugeordnet. Diese „Aufenthaltsorte“ können z. B. Funkzellen, momentane Endgeräteadressen, Aufenthaltsbereiche (*location areas*), Internet-Subnetze oder zuständige Teilnehmervermittlungsstellen sein². Der stochastische Prozeß bildet in einem Zufallsexperiment ξ_j jedes Element t des kontinuierlichen Parameterraums T auf ein Element $X(t, \xi_j)$ des Zustandsraums S ab. Abbildung 5-1 zeigt beispielhaft eine Realisierung $X(t, \xi_j)$ als das Ergebnis eines beliebigen Zufallsexperimentes ξ_j . Zu den Zeitpunkten T_n finden Zustandsübergänge statt. Wir definieren $X(T_n, \xi_j) = \lim_{\Delta t \rightarrow 0^+} X(T_n + \Delta t, \xi_j)$.

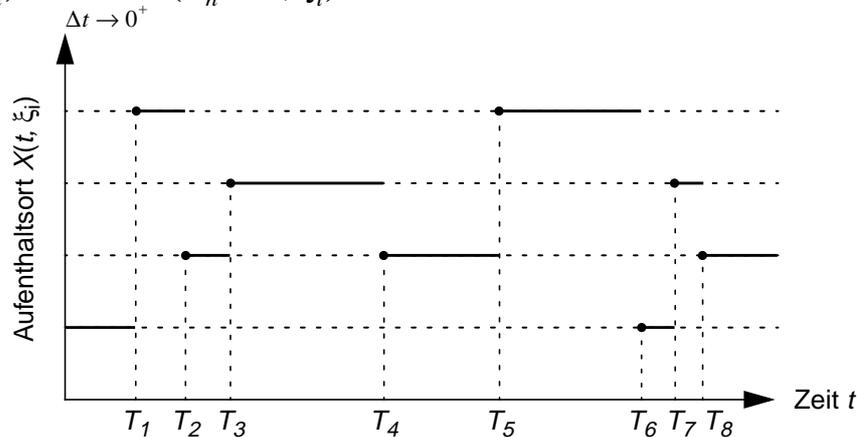


Abbildung 5-1: Bewegung als diskreter stochastischer Prozeß.

Aufgabe eines Mobilitätsmodells ist es nun, das durch den stochastischen Prozeß beschriebene Verhalten einer mobilen Instanz nachzubilden. Dazu werden häufig [z. B.: Law95, AH95, KCU95, Col95, Kwi96, MH98] Markoffsche oder Semi-Markoffsche Zustandsprozesse verwendet. Ein Zustandsprozeß sei ein stochastischer Prozeß mit abzählbarem Zustandsraum. Ist er kontinuierlich und besitzt er die Markoff-Eigenschaft³, so wird er als Markoffscher Zustandsprozeß bezeichnet. Bei einem Semi-Markoffschen Zustandsprozeß [Kü96, TG97, Gae77, KS76] hängen bei Erreichen eines Zustands u Zeitpunkt und Ziel des nächsten Zustandsübergangs nicht von früheren Zuständen oder den Verweilzeiten in denselben ab. Man kann

2. Ein diskreter stochastischer Prozeß kann einen abzählbar unendlichen Zustandsraum besitzen. Bei den hier beschriebenen Beispielen kann man jedoch ohne Einschränkung der Allgemeinheit von einem endlichen Zustandsraum ausgehen.
3. Die Markoff-Eigenschaft bedeutet, daß zu jedem beliebigen Zeitpunkt t_n aus T die zukünftige Entwicklung eines stochastischen Prozesses (für $t > t_n$) nicht vom Verlauf des Prozesses vor dem Zeitpunkt t_n abhängt. Man spricht auch davon, daß der stochastische Prozeß „gedächtnislos“ ist oder daß das gesamte Gedächtnis im aktuellen Zustand enthalten ist.

also sagen, daß ein Semi-Markoff-Prozeß zu den Zeitpunkten T_n sein Gedächtnis verliert. Die aufeinanderfolgenden Zustände $X_n = X(T_n)$ bilden damit eine sogenannte eingebettete Markoff-Kette.

Im weiteren werden nur *homogene* Prozesse betrachtet, also Prozesse, deren Charakteristik zeitlich konstant ist. Für den Markoffschen Zustandsprozeß bedeutet dies, daß die Wahrscheinlichkeit, daß der Prozeß ausgehend von einem Zustand u zur Zeit s sich nach der Zeit t im Zustand v befindet, unabhängig von s ist, daß also

$$P \left\{ X(t+s) = v \mid X(s) = u \right\} = P \left\{ X(t) = v \mid X(0) = u \right\} \quad (5-1)$$

für alle $u, v \in S$ und alle $s, t \geq 0$.

Für den Semi-Markoff-Prozeß bedeutet dies: Die Wahrscheinlichkeit, daß der Prozeß ausgehend von einem Zustandsübergang nach $X_n = u$ den nächsten Zustandsübergang innerhalb des Zeitraums t erfährt und daß dieser nach v führt, ist unabhängig von n , daß also

$$\begin{aligned} & P \left\{ X_{n+1} = v, T_{n+1} - T_n \leq t \mid X_n = u \right\} \\ &= P \left\{ X_{m+1} = v, T_{m+1} - T_m \leq t \mid X_m = u \right\} \end{aligned} \quad (5-2)$$

für alle $u, v \in S$, alle $t \geq 0$ und alle m, n .

Beim Markoffschen Zustandsprozeß müssen die Verweildauern in den einzelnen Zuständen negativ-exponentiell verteilt und unabhängig vom nachfolgenden Zustand sein. Dies ist eine direkte Folge der Gedächtnislosigkeit des Prozesses. Bei einem homogenen Semi-Markoff-Prozeß dagegen können nicht nur die Verweildauern in den einzelnen Zuständen beliebig verteilt sein, sie können sich auch nach Maßgabe dessen, wohin der nachfolgende Übergang führt, unterscheiden. In einer Simulation läßt sich dieses Verhalten nachbilden, indem man beim Übergang in einen Zustand u anhand der Verteilung $p_{u,v}$ den nachfolgenden Zustand⁴ v auswürfelt und dann den Zeitraum bis zum Übergang in den Zustand v mittels einer Zufallsvariablen mit der Verteilungsfunktion $F_{u,v}(t)$ bestimmt. Zur Beschreibung dieser Größen definieren wir die bedingte Wahrscheinlichkeit

$$Q_{u,v}(t) = P \left\{ X_{n+1} = v, T_{n+1} - T_n \leq t \mid X_n = u \right\} \quad (5-3),$$

die nach (5-2) unabhängig von n ist.

4. Übergänge finden hier nur zwischen nicht-identischen Zuständen statt. Folglich gilt für alle u : $p_{u,u} = 0$.

Damit lassen sich $p_{u,v}$ und $F_{u,v}(t)$ folgendermaßen schreiben:

$$p_{u,v} = \lim_{t \rightarrow \infty} Q_{u,v}(t) \tag{5-4}$$

$$F_{u,v}(t) = \frac{Q_{u,v}(t)}{p_{u,v}} \tag{5-5}$$

Die zu $F_{u,v}(t)$ gehörige Verteilungsdichtefunktion $f_{u,v}(t)$ lautet entsprechend:

$$f_{u,v}(t) = \frac{1}{p_{u,v}} \cdot \frac{d}{dt} Q_{u,v}(t) \tag{5-6}$$

Alle $p_{u,v}$ bilden die sogenannte Zustandsübergangsmatrix **P**.

Der Markoffsche Zustandsprozeß ist als Spezialfall im Semi-Markoff-Prozeß enthalten. Die $Q_{u,v}(t)$ haben die Form $Q_{u,v}(t) = p_{u,v} \cdot (1 - e^{-\mu_u \cdot t})$, wobei $1/\mu_u$ die mittlere Verweildauer im Zustand u ist. Die $F_{u,v}(t)$ sind in diesem Fall unabhängig von v und haben folgende Form: $F_{u,v}(t) = 1 - e^{-\mu_u \cdot t}$.

Die beschriebenen Zustandsprozesse sollen dazu verwendet werden, in einfacher Weise Zustandsübergangsereignisse zu generieren, die auf für die Mobilitätsverwaltung relevante Zustandsübergänge (Aufenthaltsortswechsel) abgebildet werden können. Durch eine zweistufige Modellierung läßt sich eine Trennung von Zuständen zur Beschreibung von physikalischen Aufenthaltsorten und Zuständen zur einfachen Modellierung des Bewegungsverhaltens erreichen. Abbildung 5-2 illustriert die Abbildung von Übergängen im erzeugenden Zustandsprozeß auf Übergänge zwischen Aufenthaltsorten. Dabei wird der Zustandsprozeß durch einen orientierten Zustandsübergangsgraphen repräsentiert. Die Zustände werden als Knoten und die möglichen Zustandsübergänge als gerichtete Kanten (Pfeile) zwischen den entsprechenden Knoten dargestellt.⁵ Gruppen von Zuständen werden Aufenthaltsorte zugeordnet, und Zustandsübergänge zwischen den Gruppen werden als Aufenthaltsortswechsel interpretiert.

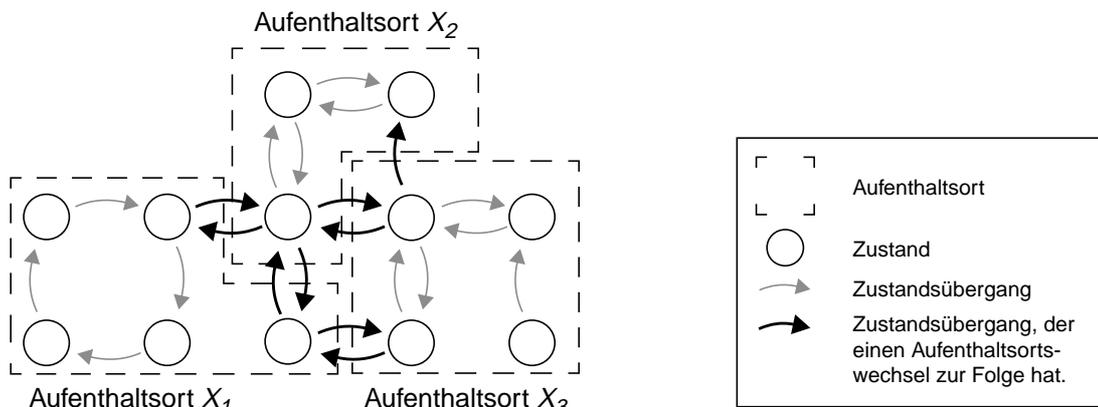


Abbildung 5-2: Modellierung der Bewegung einer mobilen Instanz mit Hilfe eines Zustandsprozesses und Abbildung von Zustandsübergängen im Zustandsübergangsgraphen auf Aufenthaltsortswchsel.

5. In weiteren Darstellungen werden gegebenenfalls die Pfeile mit den bedingten Übergangswahrscheinlichkeiten $p_{u,v}$ gewichtet.

5.1.1.2 Klassifizierung von Zuständen

Nach [Gae77] läßt sich der Zustandsraum in verschiedene Klassen einteilen. Zustand u heißt *erreichbar* von Zustand v , wenn es im gerichteten Graphen einen Weg von v nach u gibt. Dabei können beliebig viele Knoten auf dem Weg zwischen v und u liegen. Zwei verschiedene Zustände u und v heißen *verbunden*, wenn u von v und v von u erreichbar ist. Jeder Zustand ist mit sich selbst verbunden. Sind u und v sowie v und w verbunden, so sind es auch u und w . Alle Zustände, die *verbunden* sind, bilden eine sogenannte *Äquivalenzklasse* hinsichtlich der Relation „*verbunden*“. Eine solche Äquivalenzklasse wird als *irreduzibel* bezeichnet. Sind alle Zustände des Zustandsraums *verbunden*, so heißt der Zustandsprozeß ebenfalls *irreduzibel*. Ist dies nicht der Fall, so zerfällt er in mehrere Äquivalenzklassen. Eine Klasse heißt *wesentlich*, wenn sie abgeschlossen ist, d.h. wenn von ihr keine Zustände außerhalb der Klasse erreicht werden können. Ansonsten heißt sie *nicht-wesentlich*. Entsprechend ihrer Klassenzugehörigkeit werden die Zustände in *wesentliche* und *nicht-wesentliche* Zustände unterteilt.

Bei einem endlichen Zustandsraum stimmen die Begriffe *wesentlich* und *nicht-wesentlich* mit den Begriffen *positiv-rekurrent* und *transient* überein [Gae77]. Ein Zustand u heißt *rekurrent*, wenn der Prozeß ausgehend von u in u verbleibt (*absorbierender Zustand*) oder nach Verlassen von u mit Wahrscheinlichkeit 1 nach u zurückkehrt. Er heißt *positiv-rekurrent*, wenn die Rückkehr in endlicher Zeit erfolgt.⁶ Ist ein Zustand nicht rekurrent, so heißt er *transient*. Der Prozeß verbleibt nicht in transienten Zuständen und die Rückkehrwahrscheinlichkeit nach Verlassen ist kleiner als 1. Unabhängig vom Ausgangszustand verläßt ein Prozeß mit endlichem Zustandsraum im Lauf der Zeit die Menge der transienten Zustände und verbleibt danach in der Menge der rekurrenten Zustände. Ist die Menge der rekurrenten Zustände irreduzibel, so ergibt sich nach [Gae77] unabhängig vom Ausgangszustand für $t \rightarrow \infty$ für jeden rekurrenten Zustand eine bestimmte Zustandswahrscheinlichkeit π_u mit

$$\sum_{u \in S} \pi_u = 1 \quad (5-7).$$

Der Prozeßverlauf kann damit in zwei Phasen eingeteilt werden: in eine transiente Phase und in eine stationäre Phase. Während sich der Prozeß in transienten Zuständen befindet, befindet er sich noch in der transienten Phase. Mit dem Eintritt in die Menge der rekurrenten Zustände beginnt er, der stationären Phase zuzustreben. Der Übergang zwischen transienter und stationärer Phase ist jedoch nicht exakt definiert. Ein stationärer stochastischer Prozeß [CM65] ist dadurch charakterisiert, daß die den Prozeß charakterisierende n -dimensionale ($n \in \mathbf{N}$) Verteilungsfunktion

$$F_{X_1 \dots X_n}(x_1, \dots, x_n, t_1, \dots, t_n) = P\{X(t_1) \leq x_1, \dots, X(t_n) \leq x_n\} \quad (5-8)$$

6. Ansonsten heißt er *null-rekurrent*. Bei einem endlichen Zustandsraum existieren jedoch keine solchen Zustände.

zeitinvariant ist, daß also für alle h und für alle t_1, \dots, t_n gilt:

$$F_{X_1 \dots X_n}(x_1, \dots, x_n, t_1, \dots, t_n) = F_{X_1 \dots X_n}(x_1, \dots, x_n, t_1 + h, \dots, t_n + h) \quad (5-9).$$

Für einen homogenen Zustandsprozeß, bei dem die Menge der rekurrenten Zustände irreduzibel ist, bedeutet dies, daß in der stationären Phase der Prozeßverlauf unabhängig vom Anfangszustand ist, und daß die Zustandswahrscheinlichkeiten π_u angetroffen werden.

5.1.2 Stationäres Modell

5.1.2.1 Die charakteristischen Größen des Zustandsprozesses

Im stationären Modell werden nur rekurrente Zustände betrachtet. Soll das Modell nicht in unabhängige Teilmodelle zerfallen, so muß gefordert werden, daß der Zustandsprozeß irreduzibel ist. Im weiteren werden nur irreduzible Zustandsprozesse mit endlichem Zustandsraum betrachtet.

Die mittlere Verweildauer $1/\mu_{u,v}$ im Zustand u vor dem Übergang in den Zustand v berechnet sich aus

$$1/\mu_{u,v} = \int_0^{\infty} \tau \cdot f_{u,v}(\tau) d\tau \quad (5-10).$$

Die mittlere Verweildauer $1/\mu_u$ im Zustand u ergibt sich dann entsprechend zu

$$1/\mu_u = \sum_{v \in S} p_{u,v} \cdot 1/\mu_{u,v} \quad (5-11),$$

wobei für alle u gilt:

$$p_{u,u} = 0 \quad \text{und} \quad \sum_{v \in S} p_{u,v} = 1 \quad (5-12).$$

Wir definieren nun die globale stationäre Zustandsübergangsrate $\phi_{u,v}$ zwischen den Zuständen u und v , und nennen diese den „Fluß“ zwischen u und v :

$$\phi_{u,v} = \pi_u \cdot \mu_u \cdot p_{u,v} \quad (5-13).$$

Den gesamten Fluß aus u heraus nennen wir ϕ_u , mit

$$\phi_u = \sum_{v \in S} \phi_{u,v} = \sum_{v \in S} \pi_u \cdot \mu_u \cdot p_{u,v} = \pi_u \cdot \mu_u \quad (5-14).$$

Entsprechend den Gleichungen des statistischen Gleichgewichts [Kü96] für stationäre Zustandsprozesse läßt sich ein statistisches Flußgleichgewicht für den Zustand u formulieren:

$$\sum_{v \in S} \phi_{u,v} = \sum_{v \in S} \phi_{v,u} \quad (5-15),$$

welches sich umformen läßt zu:

$$\pi_u \cdot \mu_u = \sum_{v \in S} \pi_v \cdot \mu_v \cdot p_{v,u} \Leftrightarrow \phi_u = \sum_{v \in S} \phi_v \cdot p_{v,u} \quad (5-16).$$

Für alle N Zustände ergibt sich damit in Vektorschreibweise:

$$\underline{\phi} = \mathbf{P}^T \cdot \underline{\phi} \quad \text{mit} \quad \underline{\phi} = \begin{bmatrix} \phi_1 \\ \dots \\ \phi_N \end{bmatrix} \quad \text{und} \quad \mathbf{P}^T = \begin{bmatrix} 0 & p_{2,1} & \dots & p_{N,1} \\ p_{1,2} & 0 & \dots & \dots \\ \dots & \dots & 0 & p_{N,N-1} \\ p_{1,N} & \dots & p_{N-1,N} & 0 \end{bmatrix} \quad (5-17).$$

5.1.2.2 Eigenschaften der Matrix \mathbf{P}^T

\mathbf{P}^T ist eine transponierte stochastische Matrix, d.h. eine Matrix mit nicht-negativen Elementen, deren Spaltensummen alle 1 ergeben. $\underline{\phi}$ ist nach Gleichung (5-17) ein Eigenvektor von \mathbf{P}^T zum Eigenwert 1. Beschreibt die Zustandsübergangsmatrix \mathbf{P} einen irreduziblen Zustandsprozeß, so existiert für \mathbf{P}^T genau ein Eigenwert zum Wert 1 und sämtliche Elemente des zugehörigen Eigenvektors sind von Null verschieden und haben dasselbe Vorzeichen, d. h. können alle positiv gemacht werden. Diese Eigenschaften von \mathbf{P}^T lassen sich aus den Eigenschaften von \mathbf{P} herleiten, was im Folgenden kurz skizziert wird.

Eine Matrix \mathbf{M} heißt nach [HLNS78] *unzerlegbar*, wenn \mathbf{M} nicht durch Umnummerierung der Zustände auf die Form:

$$\mathbf{M}^* = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{0} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \quad (5-18)$$

gebracht werden kann, wobei \mathbf{M}_{11} und \mathbf{M}_{22} quadratische Matrizen sind und $\mathbf{0}$ eine Nullmatrix ist.

Die Eigenschaft *unzerlegbar* ist invariant gegenüber Transponierung.⁷ Da \mathbf{P} unzerlegbar⁸ und nicht-negativ ist, ist es also auch \mathbf{P}^T . In [HLNS78] werden folgende Eigenschaften für eine unzerlegbare nicht-negative ($n \times n$)-Matrix \mathbf{M} aufgeführt:

- \mathbf{M} hat einen maximalen Eigenwert λ_{\max} , der betragsmäßig größergleich ist als alle anderen Eigenwerte ($\lambda_{\max} \geq |\lambda_i|$) und für den gilt $\lambda_{\max} \in \mathbf{R}$ und $\lambda_{\max} > 0$.
- λ_{\max} liegt zwischen der kleinsten und der größten Zeilensumme von \mathbf{M} .

7. Man kann leicht sehen, daß eine Matrix, die nicht auf die Form \mathbf{M}^* gebracht werden kann, auch nicht auf die Form $(\mathbf{M}^*)^T$ gebracht werden kann.

8. Könnte \mathbf{P} auf die Form nach Gleichung (5-18) gebracht werden, so wäre der Zustandsprozeß nicht mehr irreduzibel, da die zu \mathbf{M}_{22} gehörigen Zustände nicht von den zu \mathbf{M}_{11} gehörigen Zuständen erreicht werden könnten.

- Zu λ_{\max} gibt es genau einen (linear unabhängigen) Eigenvektor⁹ und dessen sämtliche Elemente sind von Null verschieden und haben dasselbe Vorzeichen.

Für die stochastische Matrix \mathbf{P} folgt aus der zweiten Eigenschaft, daß $\lambda_{\max} = 1$ ist.¹⁰ Da wegen

$$\det(\lambda\mathbf{I} - \mathbf{P}) = \det((\lambda\mathbf{I} - \mathbf{P})^T) = \det(\lambda\mathbf{I} - \mathbf{P}^T) \quad (5-19)$$

die charakteristischen Polynome zur Bestimmung der Eigenwerte von \mathbf{P} und von \mathbf{P}^T identisch sind, besitzt \mathbf{P}^T dieselben Eigenwerte wie \mathbf{P} . Folglich hat auch \mathbf{P}^T den maximalen Eigenwert $\lambda_{\max} = 1$. Also besitzt auch \mathbf{P}^T zum einfachen Eigenwert $\lambda_{\max} = 1$ einen Eigenvektor, dessen sämtliche Elemente von Null verschieden sind und dasselbe Vorzeichen haben.

Umgekehrt kann man zeigen, daß eine stochastische Matrix \mathbf{P} genau dann unzerlegbar ist, wenn ihr maximaler Eigenwert 1 eine einfache Nullstelle des charakteristischen Polynoms ist und wenn zum Eigenwert 1 der Matrix \mathbf{P}^T ein positiver Eigenvektor existiert [HLNS78]. Da eine Markoffkette genau dann irreduzibel ist, wenn die zugehörige Matrix \mathbf{P} unzerlegbar ist, kann durch Prüfung, ob 1 eine einfache Nullstelle des charakteristischen Polynoms von \mathbf{P} (bzw. von \mathbf{P}^T) ist und ob zum Eigenwert 1 von \mathbf{P}^T ein positiver Eigenvektor existiert, festgestellt werden, ob die zugehörige Markoffkette irreduzibel – und damit auch positiv-rekurrent – ist. Dies läßt sich jedoch auch algorithmisch feststellen, indem man am Zustandsübergangsgraphen überprüft, ob alle Zustände miteinander verbunden sind, ob also jeder Zustand von jedem anderen aus erreichbar ist.¹¹

5.1.3 Parametrisierung des stationären Modells

5.1.3.1 Motivation

Die charakteristischen Größen $p_{u,v}$, μ_u und π_u (bzw. in Matrix-/Vektorschreibweise: \mathbf{P} , $\underline{\mu}$ und $\underline{\pi}$) des Zustandsprozesses sind im allgemeinen nicht alle von vornherein gegeben. Darüberhinaus sind sie nicht unabhängig voneinander, sondern den Gleichungen (5-12), (5-14) und (5-17) unterworfen. Ist ein Teil dieser Parameter durch Messungen von realem Verhalten bekannt oder durch Vorgaben im Rahmen von abstrakten Parameterstudien vorgegeben, so stellt sich die Aufgabe, den verbleibenden Teil der Parameter so zu bestimmen, daß die oben referenzierten Gleichungen erhalten bleiben. Im Folgenden werden je nach Vorgabe verschiedene Vorgehensweisen zur Bestimmung der verbleibenden Parameter präsentiert. Zur Veranschaulichung wird die Beschreibung der verschiedenen Vorgehensweisen durch die Behandlung eines einfachen Modellierungsbeispiels ergänzt, das die Bewegung von Personen zwischen vier Büros, welche über zwei Gangsegmente verbunden sind, wiedergeben soll.

9. λ_{\max} ist eine einfache Nullstelle des charakteristischen Polynoms.

10. Der zugehörige Eigenvektor ist $\underline{e}=(1, 1, \dots, 1)^T$, da $\mathbf{P}\underline{e} = \underline{e}$ eine stochastische Matrix charakterisiert.

11. Siehe dazu die Definition von *erreichbar* und *verbunden* auf Seite 110.

Abbildung 5-3 zeigt die Anordnung der Büroräume und den Zustandsübergangsgraphen eines möglichen Modells für die Bewegung von Personen innerhalb dieses Komplexes. Im Anhang A.1 wird ein alternatives Modell präsentiert.

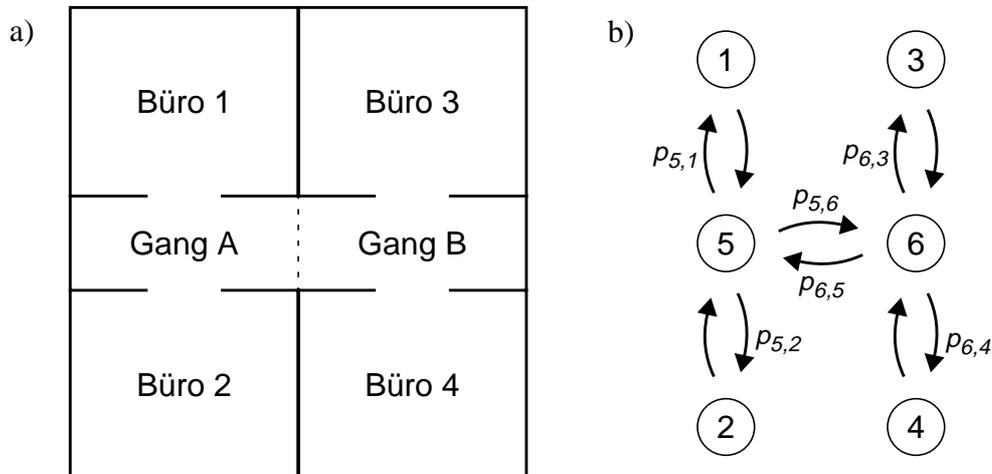


Abbildung 5-3: a) Anordnung von Büroräumen und Gangsegmenten des in diesem Kapitel betrachteten Modellierungsbeispiels sowie b) Zustandsübergangsgraph eines möglichen Modells für die Bewegung von Personen zwischen den Büroräumen.

Für die Semi-Markoff-Prozesse sind darüberhinaus noch weitere charakteristische Größen festzulegen. Dies sind die Verteilungsfunktionen $F_{u,v}(t)$ der Aufenthaltsdauern in den einzelnen Zuständen u vor einem Übergang in einen Zustand v . Wie diese sinnvoll festgelegt werden, soll hier nicht näher behandelt werden. Es wird lediglich darauf hingewiesen, daß die Gleichungen (5-11) erfüllt sein müssen, daß also μ_u , $p_{u,v}$ und die mittleren Aufenthaltsdauern $1/\mu_{u,v}$ voneinander abhängig sind.

5.1.3.2 Vorgabe der $p_{u,v}$

Der einfachste Fall ergibt sich, wenn die Zustandsübergangsmatrix \mathbf{P} vorgegeben wird. Der Eigenvektor $\underline{\phi}$ von \mathbf{P}^T zum Eigenwert 1 kann aus $\underline{\phi} = \mathbf{P}^T \cdot \underline{\phi}$ bestimmt werden. Unter Verwendung von $\phi_u = \pi_u \cdot \mu_u$ können nun noch die π_u oder die μ_u vorgegeben werden. Da der Eigenvektor $\underline{\phi}$ mit einem beliebigen positiven Faktor skaliert werden kann, verbleibt noch ein Freiheitsgrad für die Bestimmung seiner Elemente. Gibt man $\underline{\pi}$ vor, so kann der Freiheitsgrad dazu verwendet werden, eine globale Zustandswechselrate $\sum \phi$ vorzugeben, so daß gilt:

$$\sum \phi_u = \sum \phi \quad (5-20).$$

Gibt man dagegen für jeden Zustand entweder ein π_u oder ein μ_u vor, so muß der Freiheitsgrad zur Sicherstellung von $\sum \pi_u = 1$, also von:

$$\sum_{u \in A} \pi_u + \sum_{u \in S-A} \phi_u / \mu_u = 1 \quad (5-21)$$

verwendet werden. Dabei sei A die Menge aller Zustände, für die ein π_u vorgegeben ist.

Zur Illustration werden für das Beispiel von Abbildung 5-3 alle $p_{u,v}$, die nicht von vornherein den Wert 0 oder 1 haben, unter Einhaltung von $\sum_{v \in S} p_{u,v} = 1$ frei vorgegeben und zwar zu:

$$p_{5,1} = p_{5,2} = p_{6,3} = p_{6,4} = \frac{1}{3} \text{ und } p_{6,5} = p_{5,6} = \frac{1}{3}.$$

Die Lösung des zugehörigen Gleichungssystems ergibt:

$$\phi_1 = \phi_2 = \phi_3 = \phi_4 = 4 \cdot \phi_5 = 4 \cdot \phi_6 = \frac{1}{12} \cdot \sum \phi$$

Aus Vorgabe von:

$$\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0,24 \text{ und } \mu_5 = \mu_6 = \frac{1}{10 \text{ Sek.}}$$

folgt:

$$\pi_5 = \pi_6 = 0,02 \text{ und } \mu_1 = \mu_2 = \mu_3 = \mu_4 = \frac{1}{480 \text{ Sek.}} = \frac{1}{6 \text{ Min.}}.$$

5.1.3.3 Vorgabe von Gewichten zur Bestimmung der $p_{u,v}$

Um eine sinnvolle Vorgabe der $p_{u,v}$ zu unterstützen, wird hier eine einfache Regel zu deren Festlegung vorgeschlagen. Sie basiert auf der Vergabe von beliebigen Gewichtswerten G_u ($G_u > 0$) an alle Zustände des Zustandsprozesses und erzeugt eine transponierte Zustandsübergangsmatrix, deren Eigenvektor zum Eigenwert 1 unmittelbar angegeben werden kann. Darüberhinaus garantiert sie, daß die Grundvoraussetzung für die Zeitreversibilität des Zustandsprozesses:

$$\phi_{u,v} = \phi_{v,u} \text{ für alle } (u,v) \in S^2 \quad (5-22)$$

gegeben ist. Zeitreversibilität bedeutet, daß das Verhalten des Zustandprozesses nicht davon abhängt, in welcher Richtung der eindimensionale Parameterraum T durchlaufen wird. Für Markoffsche Zustandsprozesse ist Voraussetzung (5-22) hinreichend bezüglich Zeitreversibilität. Für Semi-Markoffsche Zustandsprozesse muß darüberhinaus noch gelten, daß für alle $F_{u,v}(t)$ unabhängig von v gilt: $F_{u,v}(t) = F_u(t)$.

Die Regel zur Erzeugung der $p_{u,v}$ wird im Folgenden beschrieben. Zuerst wird festgelegt, zwischen welchen Zuständen überhaupt Übergänge zulässig sind. Dazu definieren wir die Nachbarschaftsmatrix \mathbf{N} mit den Elementen $n_{u,v}$, für die gilt:

$$n_{u,v} = \begin{cases} 0 & \text{falls } u = v \\ 0 & \text{falls ein Übergang von } u \text{ nach } v \text{ nicht zulässig ist} \\ 1 & \text{falls ein Übergang von } u \text{ nach } v \text{ zulässig ist} \end{cases} \quad (5-23).$$

Ausgehend von einer symmetrischen Nachbarschaftsmatrix werden die Zustandsübergangswahrscheinlichkeiten dann zu

$$p_{u,v} = \frac{G_v \cdot n_{u,v}}{\sum_{w \in S} G_w \cdot n_{u,w}} \quad (5-24)$$

gesetzt. Der zugehörige Zustandsprozeß ist zeitreversibel und die zugehörigen ϕ_u ergeben sich zu:

$$\phi_u = G_u \cdot \sum_{w \in S} (G_w \cdot n_{u,w}) \quad (5-25),$$

woraus schließlich folgt, daß

$$\phi_{u,v} = G_u \cdot G_v \cdot n_{u,v} \quad (5-26).$$

Dies wird im Folgenden gezeigt: Setzt man (5-24) in die Gleichung für das statistische Flußgleichgewicht (5-16) ein, so ergibt sich

$$\phi_u = \sum_{v \in S} \phi_v \cdot \frac{G_u \cdot n_{v,u}}{\sum_{w \in S} G_w \cdot n_{v,w}} \Leftrightarrow \frac{\phi_u}{G_u} = \sum_{v \in S} \frac{\phi_v \cdot n_{v,u}}{\sum_{w \in S} G_w \cdot n_{v,w}} \quad (5-27).$$

Mit Hilfe der Substitution

$$\phi_u^* = \frac{\phi_u}{G_u \cdot \sum_{w \in S} G_w \cdot n_{u,w}} \quad (5-28)$$

erhält man:

$$\phi_u^* \cdot \sum_{w \in S} G_w \cdot n_{u,w} = \sum_{v \in S} \phi_v^* \cdot G_v \cdot n_{v,u} \quad (5-29),$$

was für alle Zustände in die Form

$$\underline{\phi}^* = \mathbf{G} \cdot \underline{\phi}^* \quad (5-30)$$

gebracht werden kann, wobei die Matrix \mathbf{G} aus den Elementen

$$g_{u,v} = \frac{G_v \cdot n_{v,u}}{\sum_{w \in S} G_w \cdot n_{u,w}} \quad (5-31)$$

besteht. Die Zeilensummen von \mathbf{G} ergeben unter Verwendung der Symmetrie der Nachbarschaftsmatrix \mathbf{N} ($n_{u,v} = n_{v,u}$):

$$\sum_{v \in S} g_{u,v} = \frac{\sum_{v \in S} G_v \cdot n_{u,v}}{\sum_{w \in S} G_w \cdot n_{u,w}} = 1 \quad (5-32),$$

was bedeutet, daß \mathbf{G} eine (unzerlegbare) stochastische Matrix ist, und folglich

$$\underline{\phi}^* = \underline{e} = (1, 1, \dots, 1)^T \quad (5-33)$$

einzigem (linear unabhängiger) Eigenvektor von \mathbf{G} zum Eigenwert 1 und damit einzige Lösung von (5-30) ist. Daraus folgt mit Hilfe von Gleichung (5-28) sofort Gleichung (5-25) – *q. e. d.*

Entsprechend läßt sich die Bedingung für Zeitreversibilität umformen:

$$\begin{aligned}
 \phi_u \cdot p_{u,v} &= \phi_v \cdot p_{v,u} \\
 &\Leftrightarrow \\
 \phi_u \cdot \frac{G_v \cdot n_{u,v}}{\sum_{w \in S} G_w \cdot n_{u,w}} &= \phi_v \cdot \frac{G_u \cdot n_{v,u}}{\sum_{w \in S} G_w \cdot n_{v,w}} \\
 &\Leftrightarrow \\
 \phi_u^* \cdot n_{u,v} &= \phi_v^* \cdot n_{v,u}
 \end{aligned} \tag{5-34}$$

was aufgrund von Gleichung (5-33) stets erfüllt ist.

Im Beispiel nach Abbildung 5-3 ergeben sich die freien $p_{u,v}$ zu $1/3$, wenn alle Gewichte G_u identisch gewählt werden – also dieselbe Lösung wie in der Rechnung auf Seite 115.

5.1.3.4 Vorgabe der ϕ_u

Der schwierigere Fall liegt vor, wenn nicht \mathbf{P} , sondern sowohl $\underline{\pi}$ als auch $\underline{\mu}$ – und damit $\underline{\phi}$ – vorgegeben werden. Abhängig von der Wahl der Vorgaben und der Zahl der zu bestimmenden $p_{u,v}$ ist das Gleichungssystem $\underline{\phi} = \mathbf{P}^T \cdot \underline{\phi}$ entweder eindeutig lösbar, unterbestimmt oder nicht lösbar.

Im Beispiel von Abbildung 5-3 etwa sind die sechs unbekanntes $p_{u,v}$ mit Hilfe der sechs Gleichungen des Gleichungssystems $\underline{\phi} = \mathbf{P}^T \cdot \underline{\phi}$ und den Randbedingungen $p_{u,v} > 0$ und $\sum_{v \in S} p_{u,v} = 1$ entweder eindeutig festgelegt, oder nicht bestimmbar. Sie ergeben sich zu:

$$\begin{aligned}
 p_{5,1} &= \frac{\phi_5}{\phi_1}, p_{5,2} = \frac{\phi_5}{\phi_2}, p_{6,3} = \frac{\phi_6}{\phi_3}, p_{6,4} = \frac{\phi_6}{\phi_4} \\
 p_{6,5} &= \frac{\phi_5 - \phi_1 - \phi_2}{\phi_6}, p_{5,6} = \frac{\phi_6 - \phi_3 - \phi_4}{\phi_5}
 \end{aligned} \tag{5-35}$$

unter der Bedingung, daß $p_{u,v} > 0$ sowie

$$\frac{\phi_5}{\phi_1} + \frac{\phi_5}{\phi_2} + \frac{\phi_6 - \phi_3 - \phi_4}{\phi_5} = 1 \quad \text{und} \quad \frac{\phi_6}{\phi_3} + \frac{\phi_6}{\phi_4} + \frac{\phi_5 - \phi_1 - \phi_2}{\phi_6} = 1 \quad \text{erfüllt sind.}$$

Gibt man etwa:

$$\phi_1 = \phi_2 = \phi_3 = \phi_4 = 0,24 \cdot \frac{1}{8 \text{ Min.}} = \frac{1}{2000 \text{ Sek.}} = 0,0005 \frac{1}{\text{Sek.}} \text{ und}$$

$$\phi_5 = \phi_6 = 0,02 \cdot \frac{1}{10 \text{ Sek.}} = \frac{1}{500 \text{ Sek.}} = 0,002 \frac{1}{\text{Sek.}} \text{ vor, so ergibt sich:}$$

$$p_{5,1} = p_{5,2} = p_{6,3} = p_{6,4} = 0,25 \text{ und } p_{6,5} = p_{5,6} = 0,5.$$

Dieses Beispiel stellt jedoch einen Idealfall dar. Im allgemeinen ist das Gleichungssystem $\underline{\phi} = \mathbf{P}^T \cdot \underline{\phi}$ zur Ermittlung der $p_{u,v}$ unterbestimmt und es ist nicht einfach, weitere geeignete Gleichungen zur Festlegung der $p_{u,v}$ zu finden, welche die notwendigen Randbedingungen erfüllen. Im Folgenden wird deshalb ein einfaches Verfahren vorgeschlagen, das bei symmetrischer Nachbarschaftsmatrix mit Hilfe der im letzten Abschnitt definierten Gewichte eine spezielle Lösung für die $p_{u,v}$ ermittelt.

5.1.3.5 Vorgabe der ϕ_u und Bestimmung der $p_{u,v}$ über die Gewichte G_u

Unter Vorgabe einer symmetrischen Nachbarschaftsmatrix wird das Problem zur Bestimmung der $p_{u,v}$ auf das Problem der Bestimmung der Gewichte G_u reduziert. Gesucht sind die G_u , welche bei Anwendung der Gleichungen (5-24) Flüsse ϕ_u^\dagger mit

$$\phi_u^\dagger = G_u \cdot \sum_{v \in S} G_v \cdot n_{u,v} \quad (5-36)$$

erzeugen, die den vorgegebenen ϕ_u möglichst nahe kommen. Dazu soll die Funktion

$$\min \sum_{u \in S} \left(\frac{\phi_u^\dagger - \phi_u}{\phi_u} \right)^2 \quad (5-37)$$

unter der Randbedingung $G_u > 0$ optimiert werden. Es soll also die Summe der relativen Abweichungsquadrate der ϕ_u^\dagger von den ϕ_u minimiert werden. Als Startwert für die Optimierung haben sich folgende Werte als günstig erwiesen:

$$G_u^{\text{Start}} = \sqrt{\frac{\phi_u}{\sum_{v \in S} n_{u,v}}} \quad (5-38).$$

Sie beruhen auf der Annahme, daß alle Gewichtswerte G_v der Nachbarn eines Zustands u identisch mit G_u sind. Gleichung (5-38) folgt also aus der Annahme, daß:

$$\phi_u = G_u \cdot \sum_{v \in S} G_u \cdot n_{u,v}.$$

Für den Fall, daß die ϕ_u^\dagger von den ϕ_u abweichen, weichen natürlich auch die zugehörigen Zustandswahrscheinlichkeiten und/oder die zugehörigen mittleren Verweildauern von den ursprünglich angenommenen Werten ab. Je nachdem, welche der Vorgaben als wichtiger betrachtet werden, müssen die weniger wichtigen analog zum Vorgehen bei der direkten Vorgabe der $p_{u,v}$ angepaßt werden.

Bei manchen Lösungen drohen Flüsse zwischen Zuständen so klein zu werden, daß der Zustandsübergangsgraph in nahezu unabhängige Teilgraphen zerfällt. Will man dies verhindern, so bietet es sich an, minimale Flüsse $\phi_{u,v}^{\min}$ zwischen den Zuständen zu definieren. Auf Basis der Gleichung (5-26) lassen sich damit für die G_u weitere Randbedingungen formulieren:

$$G_u \cdot G_v - \phi_{u,v}^{\min} > 0 \quad \forall u \in S, v \in S, n_{u,v} > 0 \quad (5-39).$$

Für ein Optimierungsverfahren stellt die Einhaltung dieser Randbedingungen einen hohen Aufwand dar, da es sich zum einen um eine große Zahl von Randbedingungen handeln kann, und da es zum anderen nicht-lineare Randbedingungen sind. Um diesen Aufwand zu verringern, verwenden wir stattdessen für die Gewichtswerte G_u die Randbedingung $G_u > G_u^{\min}$ mit den willkürlich festgelegten

$$G_u^{\min} = \alpha \cdot G_u^{\text{Start}}, \alpha \in [0, 1) \quad (5-40),$$

wobei der globale Faktor α so festgelegt wird, daß alle Gleichungen (5-39) garantiert erfüllt sind. Dies wird durch

$$\alpha = \max \sqrt{\frac{\phi_{u,v}^{\min}}{G_u^{\text{Start}} \cdot G_v^{\text{Start}}}} \quad (5-41)$$

erreicht.

Die Wahl der $\phi_{u,v}^{\min}$ bleibt dem Modellersteller überlassen. Folgende Vorgabe mit einem frei wählenden Kopplungsfaktor γ wird vorgeschlagen:

$$\text{Vorschlag } \phi_{u,v}^{\min} = n_{u,v} \cdot \gamma \cdot \min \left(\frac{\phi_u}{\sum_{w \in S} n_{u,w}}, \frac{\phi_v}{\sum_{w \in S} n_{v,w}} \right), \gamma \in [0, 1] \quad (5-42),$$

was sicherstellt, daß wenigstens von einem der beiden benachbarten Zustände u und v mindestens das γ -fache seines gleichmäßig auf alle Nachbarn verteilten vorgegebenen Flusses zum anderen Zustand fließt. Sinnvollerweise wählt man γ deutlich kleiner als 1.

Gibt man etwa im Beispiel von Abbildung 5-3 vor, daß

$$2\phi_1 = 2\phi_2 = 2\phi_3 = 2\phi_4 = \phi_5 = \phi_6 = 0,001,$$

so ergibt sich nach Gleichung (5-35):

$$p_{5,1} = p_{5,2} = p_{6,3} = p_{6,4} = 0,5 \quad \text{und} \quad p_{6,5} = p_{5,6} = 0.$$

Diese Lösung kann mit dem hier präsentierten Verfahren nach Gleichung (5-24) nicht gefunden werden, da für $n_{u,v} = 1$ die aus den $G_u > 0$ erzeugten $p_{u,v}$ nicht zu 0 werden können.

Ein numerischer Optimierungsalgorithmus ohne begrenzende $\phi_{u,v}^{\min}$ erzeugt daraus je nach Verfahren und Genauigkeit der Zahlendarstellung ein System, bei dem

$$\phi_{\{1,2,3,4\}}^{\dagger} = \phi_{\{1,2,3,4\}} - \varepsilon \quad \text{und} \quad \phi_{\{5,6\}}^{\dagger} \approx \phi_{\{5,6\}} + 4\varepsilon$$

und bei dem nur ein sehr geringer Fluß zwischen den beiden Gangsegmenten stattfindet ($p_{5,6} \approx 0,0\%$). Wählt man dagegen $\gamma = 0,3$, so weichen die ϕ_u^{\dagger} zwar stärker von den ϕ_u ab:

$$\phi_{\{1,2,3,4\}}^{\dagger} \approx 0,967 \cdot \phi_{\{1,2,3,4\}} \quad \text{und} \quad \phi_{\{5,6\}}^{\dagger} \approx 1,067 \cdot \phi_{\{5,6\}},$$

der Fluß zwischen den Gangsegmenten jedoch bleibt besser erhalten ($p_{5,6} = p_{6,5} \approx 9,4\%$) und erfüllt damit die Randbedingungen: $\phi_{5,6}^{\dagger} \geq \gamma \cdot \frac{\phi_5}{3}$ und $\phi_{6,5}^{\dagger} \geq \gamma \cdot \frac{\phi_6}{3}$ von Gleichung (5-42).

5.1.3.6 Zusammenfassung der verschiedenen Vorgehensweisen

Abbildung 5-4 stellt zusammenfassend die verschiedenen Vorgehensweisen zur Parametrisierung des stationären Bewegungsmodells dar. Auf der linken Seite werden die $p_{u,v}$ vorgegeben, während im rechten Teil der Darstellung die $p_{u,v}$ zu den vorgegebenen μ_u und π_u (unter Berücksichtigung der Nachbarschaftsmatrix \mathbf{N}) bestimmt werden. Sind die ϕ_u im linken Teil ermittelt, so muß entweder die Summe $\sum \phi_u$ oder mindestens eines der μ_u vorgegeben werden. Die verbleibenden Freiheitsgrade können genutzt werden, um die entsprechenden π_u frei zu wählen. Die restlichen Größen sind dann nach Skalierung der ϕ_u zur Erfüllung der Gleichung (5-21) bzw. (5-20) über $\phi_u = \pi_u \cdot \mu_u$ festgelegt. Weichen die auf der rechten Seite zu den vorgegebenen π_u und μ_u ermittelten ϕ_u^{\dagger} von den entsprechenden ϕ_u ab, so muß ein Teil der vorgegebenen Größen wie bei direkter Vorgabe der $p_{u,v}$ entsprechend dem Vorgehen auf der linken Seite neu ermittelt werden.

5.1.4 Überlagerung von Bewegungsmodellen

Bei der bisherigen Behandlung des Bewegungsmodells wurde implizit von einem einzelnen Teilnehmer ausgegangen. Geht man davon aus, daß das Verhalten einzelner Teilnehmer nicht mit dem Verhalten anderer Teilnehmer gekoppelt ist, so kann man durch einfache Überlagerung der Mobilitätsmodelle die Gesamtheit der Teilnehmerbewegungen modellieren. Alle Teilnehmer, deren Bewegungsverhalten durch einen identischen Zustandsprozeß beschrieben wird, können in eine Teilnehmerklasse k zusammengefaßt werden. Geht man davon aus, daß sich innerhalb einer Klasse k genau $\Pi^{(k)}$ Teilnehmer befinden, so kann man die charakteristischen

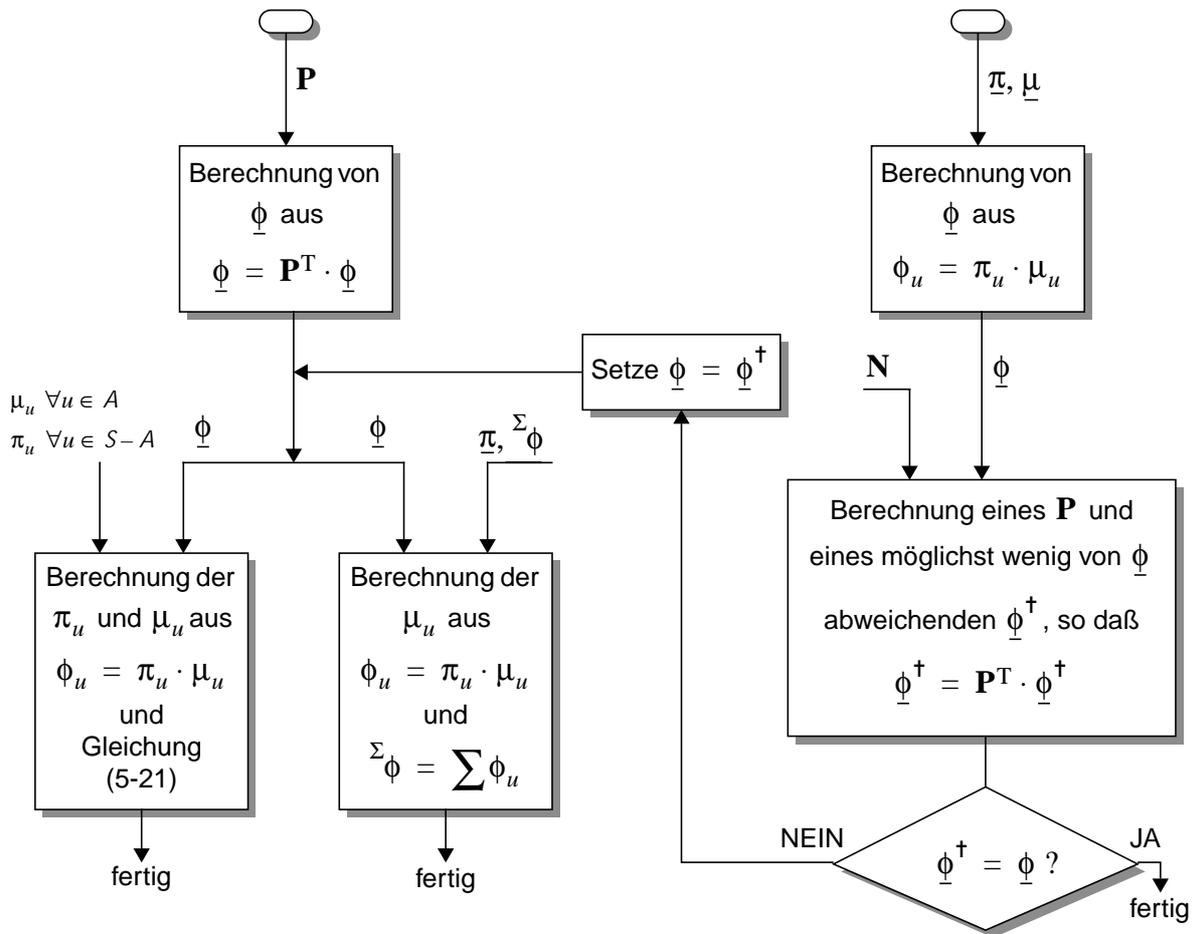


Abbildung 5-4: Verschiedene Vorgehensweisen zur Parametrisierung des stationären Bewegungsmodells je nach Art der vorgegebenen Größen.

Größen¹² des stationären Modells eines einzelnen Teilnehmers auch für den Fall mehrerer Teilnehmer interpretieren. Aus der Zustandswahrscheinlichkeit $\pi_u^{(k)}$ für einen Teilnehmer der Klasse k im Zustand u kann man die mittlere Zahl von Teilnehmern der Klasse k im Zustand u ermitteln:

$$\Pi_u^{(k)} = \Pi^{(k)} \cdot \pi_u^{(k)} \tag{5-43},$$

dabei gilt analog zu Gleichung (5-7):

$$\sum_{u \in S} \Pi_u^{(k)} = \sum_{u \in S} \Pi^{(k)} \cdot \pi_u^{(k)} = \Pi^{(k)} \tag{5-44}.$$

Entsprechend kann mit Hilfe von $\phi_{u,v}^{(k)}$ der Fluß von Teilnehmern einer Klasse k aus einem Zustand u in einen Zustand v angegeben werden:

$$\Phi_{u,v}^{(k)} = \Pi^{(k)} \cdot \phi_{u,v}^{(k)} \tag{5-45}.$$

12. Alle Größen des Bewegungsmodells werden im folgenden entsprechend ihrer Klassenzugehörigkeit zu einer Klasse k mit dem Index $^{(k)}$ versehen. Ohne Beschränkung der Allgemeinheit nehmen wir an, daß $S^{(k)}$ identisch mit S ist.

Woraus sich der Fluß von Teilnehmern einer Klasse k aus einem Zustand u zu

$$\Phi_u^{(k)} = \Pi^{(k)} \cdot \phi_u^{(k)} \quad (5-46)$$

ergibt.

Überlagert man schließlich alle Teilnehmerklassen (der Menge K), so ergibt sich für die mittlere Zahl von Teilnehmern Π_u im Zustand u , für den Teilnehmerfluß $\Phi_{u,v}$ zwischen u und v sowie für den Teilnehmerfluß Φ_u aus u heraus:

$$\Pi_u = \sum_{k \in K} \Pi_u^{(k)}, \quad \Phi_{u,v} = \sum_{k \in K} \Phi_{u,v}^{(k)}, \quad \Phi_u = \sum_{k \in K} \Phi_u^{(k)} \quad (5-47).$$

5.1.5 Kopplung des Bewegungsverhaltens und transientes Modell

Die Annahme, daß das Bewegungsverhalten eines Teilnehmers unabhängig von der Bewegung anderer Teilnehmer betrachtet werden kann, stellt eine gewisse Vereinfachung der Realität dar. Bilden etwa mehrere Teilnehmer eine Gruppe, die sich in einem öffentlichen oder privaten Verkehrsmittel befindet, so erfolgt der Aufenthaltsortswchsel der beteiligten Personen quasi gleichzeitig.¹³ Ein ähnlicher Effekt tritt auf, wenn Fahrzeug- oder Personengruppen durch Staus oder Verkehrsampeln entstehen. Neben solch einer örtlichen Kopplung des Bewegungsverhaltens kommt es häufig auch zu einer zeitlichen Kopplung. So bewegen sich zum Beispiel morgens viele Personen von der Wohnung in die Schule oder zum Arbeitsplatz. Die Zeiten, zu denen die Personen am Ziel sein wollen, können sich dabei in einem relativ kleinen Zeitfenster häufen. In ähnlicher Weise kann kulturell geprägtes Verhalten zur Kopplung des Bewegungsverhaltens vieler Teilnehmer führen. Betrachtet man auch das Ein- und Ausbuchen von Endgeräten als eine Form der Bewegung, so führt das Aus- bzw. Einschalten von Endgeräten vor bzw. nach einem Gottesdienstbesuch oder einem Mittagsgebet zu einer zeitlichen Kopplung des Bewegungsverhaltens einer großen Zahl von Teilnehmern an verschiedenen Orten.

Hier werden zwei unterschiedliche Vorgehensweisen zur Berücksichtigung der Kopplung des Teilnehmergehaltens vorgeschlagen. Das erste Verfahren ordnet im stationären, überlagerten Bewegungsmodell Zustandsübergänge nicht mehr einzelnen Teilnehmern zu, sondern Gruppen von Teilnehmern. Das zweite Verfahren beruht darauf, bei transienten Zustandsprozessen durch die Vorgabe der Startwerte eine gewisse Kopplung des Bewegungsverhaltens von Teilnehmern zu erreichen.

5.1.5.1 Zuordnung von Teilnehmergruppen zu Zustandsübergängen

Strenggenommen handelt es sich bei dieser Vorgehensweise nicht um ein Modell des Bewegungsverhaltens, sondern um eine Berücksichtigung von Gruppenbildung bei einer approximativen, analytischen Leistungsbewertung. Überlagert man die oben dargestellten, stationären

13. In wie weit dies zu quasi gleichzeitigen Aufenthaltsortsaktualisierungen führt, hängt vom Aktualisierungsverfahren ab.

Bewegungsmodelle mehrerer Klassen und einer Vielzahl von Teilnehmern, so kann man die Zwischenankunftszeit zwischen zwei Zustandswechseln am Übergang zwischen den Zuständen u und v näherungsweise als gedächtnislos und damit als negativ-exponentiell verteilt mit Mittelwert $1/\Phi_{u,v}$ betrachten. Anstelle solch eines Zufallsprozesses soll nun ein Prozeß betrachtet werden, bei dem Gruppen von Teilnehmern zwischen zwei Zuständen wechseln. Bezeichnet man die mittlere Gruppengröße als $\Gamma_{u,v}$, so kann man das stationäre Bewegungsmodell den Fluß $\Phi_{u,v}$ zwischen den Zuständen um den Faktor $1/\Gamma_{u,v}$ reduzieren. Es ergeben sich dann Gruppenzwischenankunftszeiten, die negativ-exponentiell verteilt mit Mittelwert $\Gamma_{u,v}/\Phi_{u,v}$ sind. Modelliert man die Gruppengröße als eine diskrete Zufallsvariable mit Erwartungswert $\Gamma_{u,v}$ und Verteilungsdichte $g_{u,v}(x)$, so kann man nach Bestimmung der Gruppengröße den Anteil der verschiedenen Teilnehmerklassen entsprechend dem Verhältnis der verschiedenen $\Phi_{u,v}^{(k)}$ festlegen. Greift man ein beliebiges Gruppenmitglied heraus, so ist die Wahrscheinlichkeit, daß es zur Teilnehmergruppe k , gehört $\Phi_{u,v}^{(k)}/\Phi_{u,v}$.

Bei dieser Vorgehensweise geht der Bezug zu individuellem Teilnehmerverhalten verloren. Insbesondere die Vergangenheit eines Teilnehmers bleibt verborgen, so daß es sich im wesentlichen für die Analyse von Mobilitätsverwaltungsverfahren eignet, bei denen der Protokollablauf beim Wechsel zwischen zwei Aufenthaltsbereichen unabhängig von zuvor besuchten Aufenthaltsbereichen ist.

5.1.5.2 Vorgabe der Startwerte von transienten Zustandsprozessen

Betrachtet man transiente Zustandsprozesse, die gleichzeitig gestartet werden, parallel ablaufen und gleiche oder ähnliche Parameter haben, so ergibt sich durch die Abhängigkeit des Verlaufs vom Startzustand auch eine Korrelation zwischen dem Verlauf der verschiedenen Prozesse. Dies kann beispielsweise dazu genutzt werden, um transientes, gekoppeltes Verhalten einer Vielzahl von Teilnehmern zu modellieren. Solch eine Vorgehensweise eignet sich besonders bei der simulativen Leistungsbewertung, da sich die Simulation des Bewegungsverhaltens nicht wesentlich vom stationären Fall unterscheidet und lediglich Meßmethoden der instationären Simulation verwendet werden müssen. Für die Simulationssteuerung hat dies zur Folge, daß der Ablauf ausgehend von einem vordefinierten Startzustand viele Male durchgespielt werden muß.

Zur Illustration sei ein Beispiel zur Modellierung von Teilnehmerbewegung in der morgendlichen Hauptverkehrszeit – also wenn sich viele Menschen auf dem Weg zur Arbeit befinden – dargestellt. Eine Stadt habe einen Innenbereich, zu dem aus dem Umland jeden Morgen eine Vielzahl von Pendlern strömt. Der Innenbereich sei in vier Aufenthaltsgebiete eingeteilt und von 20 Aufenthaltsgebieten des Umlandes umgeben. Abbildung 5-5 zeigt einen Zustandsübergangsgraphen, der für alle Teilnehmer identisch sei. Jeder der vielen Teilnehmer befinde sich zu Beginn in einem der Startzustände. Im Lauf der Zeit bewegen sich die Teilnehmer über die

Zwischenzustände in die Zielzustände, die als absorbierende Zustände dargestellt sind. Zustandsübergänge, bei denen Aufenthaltsbereichsgrenzen überschritten werden, bedeuten eine Bewegung eines Teilnehmers, die für die Mobilitätsverwaltung relevant ist.

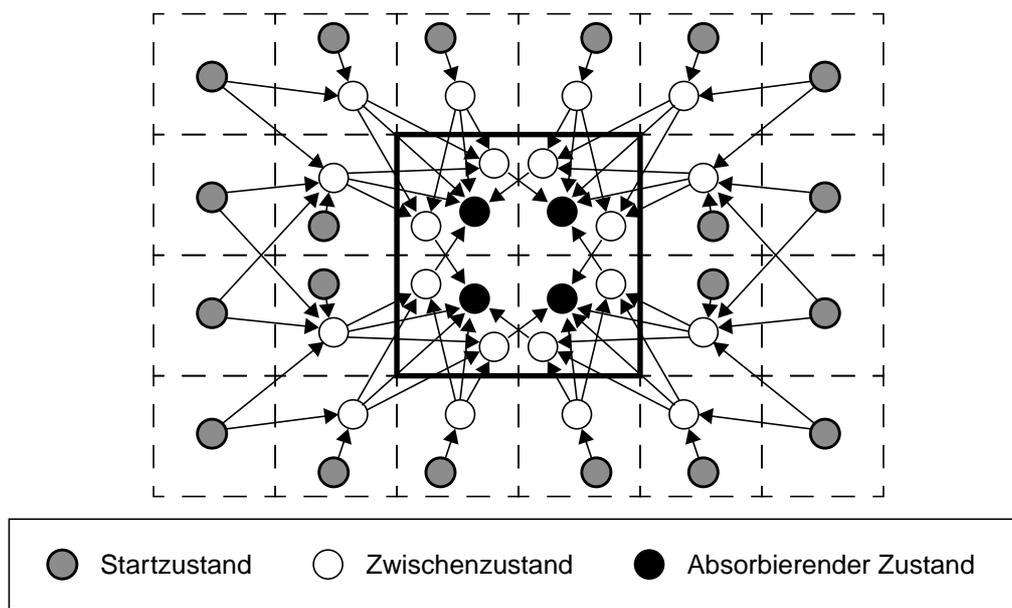


Abbildung 5-5: Modellierung von Teilnehmerbewegung zur morgendlichen Hauptverkehrszeit (*rush hour*) mit Hilfe von transienten Zustandsprozessen.

5.2 Das Teilnehmermodell: Rufe

Die Mobilitätsverwaltung in einem Kommunikationsnetz verfolgt die Bewegung mobiler Teilnehmer, um diese für Kommunikation mit anderen Teilnehmern erreichbar zu halten. Je nachdem, wie präzise und wie weit die Information über den aktuellen Aufenthaltsort eines mobilen Teilnehmers im Netz verbreitet ist, um so mehr oder weniger aufwendig ist die Ermittlung von Information, die benötigt wird, um eine Kommunikationsbeziehung zu einem mobilen Teilnehmer aufzubauen. Zur Bewertung verschiedener Verfahren der Mobilitätsverwaltung ist es daher von großer Bedeutung, wie häufig und an welchen Stellen diese Information nachgefragt wird. Der Auslöser solch einer Anfrage wird im folgenden als „Ruf“ bezeichnet. Das Konzept des Rufes beschreibt dabei den Versuch, mit Hilfe eines eindeutigen Namens oder einer quasi-permanenten Adresse eine Kommunikationsbeziehung zu einer (mobilen) Instanz aufzubauen.

In klassischen verbindungsorientierten Telekommunikationsnetzen unterscheidet man rufende und gerufene Instanz. Dadurch, daß hierbei ein Ruf zumeist mit dem Aufbau bidirektionaler Verbindungen¹⁴ verknüpft ist, muß lediglich die momentane Adresse der gerufenen Instanz ermittelt werden, während die rufende Instanz indirekt über eine Verbindungsreferenz adressiert werden kann.

14. Sowohl Nutzkanal- als auch Signalisierverbindungen sind zumeist bidirektional.

In IP-basierten Netzen werden zwar auch Verbindungen zwischen Instanzen der Transport- bzw. Anwendungsschicht aufgebaut, diese stützen sich jedoch auf die verbindungslose Kommunikation der Vermittlungsschicht („Internetschicht“) ab. Erfolgt dort die Mobilitätsunterstützung, so ist jedes einzelne Datenpaket (unabhängig von der Richtung) als „Ruf“ im obigen Sinne zu betrachten, da in jedem Datenpaket die Zielinstanz über ihre quasi-permanente Adresse adressiert wird und diese erst mit Hilfe der Mobilitätsverwaltung auf eine temporäre Adresse umgesetzt werden muß.

Während man in klassischen Telekommunikationsnetzen davon ausgehen kann, daß die Zwischenankunftszeiten zwischen Rufen unabhängig und identisch verteilt sind, läßt sich der Ankunftsprozeß von IP-Paketen nur durch sehr komplexe stochastische Prozesse beschreiben. In [KKR97] beispielsweise wird WWW-Verkehr auf sechs verschiedenen Aktivitätsebenen charakterisiert, die den Verkehr jeweils auf einer anderen Zeitskala beschreiben. Die Entwicklung eines geeigneten Modells für Internetverkehr würde den Rahmen dieser Arbeit sprengen. Im weiteren wird deshalb angenommen, daß für die Zwischenankunftszeiten von IP-Paketen, die zu einem bestimmten Ziel geschickt werden sollen, eine Verteilungsfunktion angebar ist, welche dann zum Beispiel zur Bewertung von *Caching*-Verfahren bei der Mobilitätsverwaltung verwendet werden kann.

Im Folgenden wird die Modellierung von Ursprung und Rate von Rufen näher behandelt. Dies erfolgt in Bezug auf die gerufene Instanz. Es wird also betrachtet, von wo und wie häufig ein bestimmter Teilnehmer „gerufen“ bzw. „gesucht“ wird.

5.2.1 Darstellung als Verkehrsmatrix

Obwohl rufenden Instanzen der aktuelle Aufenthaltsort einer gerufenen Instanz nicht unmittelbar bekannt ist, läßt sich annehmen, daß Ursprung und Rate von Rufen zu einem Teilnehmer von seinem aktuellen Aufenthaltsort abhängig sind. Gegebenenfalls könnte man sogar annehmen, daß nicht nur eine Abhängigkeit vom aktuell besuchten Aufenthaltsort besteht, sondern auch von der „Vergangenheit“ der Bewegung. Dies soll hier jedoch nur insofern berücksichtigt werden, als daß Ursprung und Rate von Rufen als abhängig vom aktuellen Zustand des Bewegungsmodells betrachtet werde. Damit erfolgt eine feinere Unterscheidung als auf der Ebene der Aufenthaltsorte, und die „Gedächtnislosigkeit“ des Prozesses bleibt erhalten. Ohne Einschränkung der Allgemeinheit wird im folgenden das Rufmodell auf der Basis von Zuständen des Bewegungsmodells dargestellt. Analog könnte auch eine Betrachtung auf der Basis von Aufenthaltsorten bzw. von Gruppen von Zuständen des Bewegungsmodells erfolgen.

Im weiteren gehen wir davon aus, daß Rufe zu einem mobilen Teilnehmer als ein homogener Prozeß modelliert werden können, daß also Ursprung und Rate von Rufen nur vom aktuellen Zustand im Bewegungsmodell, nicht jedoch vom Parameter *Zeit* abhängig sind. Für einen Teilnehmer der Klasse k definieren wir die Matrix $\mathbf{R}^{(k)}$, deren Elemente $r_{u,v}^{(k)}$ angeben, mit wel-

cher Rate Rufe von Teilnehmern am Ort (im Zustand) u ausgehen, wenn der gerufene Teilnehmer sich im Zustand v des Bewegungsmodells befindet:

$$\mathbf{R}^{(k)} = \begin{bmatrix} r_{u,v}^{(k)} \end{bmatrix} \quad (5-48).$$

Summiert man die $r_{u,v}^{(k)}$ über alle u auf, so ergibt sich daraus die Rate kommender Rufe zu einem Teilnehmer der Klasse k , der sich im Bewegungsmodell im Zustand v befindet:

$$r_v^{(k)} = \sum_{u \in S} r_{u,v}^{(k)} \quad (5-49).$$

Im stationären Bewegungsmodell läßt sich dann eine stationäre Verkehrsmatrix $\Lambda^{(k)}$ aufstellen, die angibt, mit welcher Rate Rufe zu einem Teilnehmer der Klasse k vom „Ort“ u zum „Ort“ v erfolgen:

$$\Lambda^{(k)} = \begin{bmatrix} \lambda_{u,v}^{(k)} \end{bmatrix} \quad (5-50),$$

wobei für die Elemente $\lambda_{u,v}^{(k)}$ von $\Lambda^{(k)}$ gilt:

$$\lambda_{u,v}^{(k)} = r_{u,v}^{(k)} \cdot \pi_v^{(k)} \quad (5-51).$$

Im folgenden Abschnitt wird ein einfaches Modell auf der Basis verschiedener „Ruftypen“ vorgestellt, das dazu dient, die Generierung der Matrizen $\mathbf{R}^{(k)}$ aus einer Reihe von anschaulichen Parametern zu ermöglichen.

5.2.2 Verwendung von Ruftypen

Folgende Anforderungen werden an ein Verfahren zur Generierung der Rufraten im Rufmodell gestellt:

- Ein Teil der Rufe soll abhängig, ein anderer Teil der Rufe soll unabhängig vom aktuellen Aufenthaltsort des gerufenen Teilnehmers sein.
- Ein Teil der Rufe soll abhängig, ein anderer Teil der Rufe soll unabhängig vom Bewegungsverhalten der rufenden Instanzen sein.
- Die globale, stationäre Ruftrate zu einem mobilen Teilnehmer soll vorgebbar sein.
- Der Einfluß von noch näher zu bezeichnenden „Distanzen“ zwischen den Zuständen soll variiert werden können.

Zur Erfüllung der ersten beiden Anforderungen unterscheiden wir vier verschiedene Ruftypen:

- **Ruftyp I:** Rufe, die *unabhängig* vom aktuellen Aufenthaltsort des gerufenen Teilnehmers erfolgen und *unabhängig* vom Bewegungsverhalten der rufenden Teilnehmer sind.
- **Ruftyp II:** Rufe, die *abhängig* vom aktuellen Aufenthaltsort des gerufenen Teilnehmers erfolgen und *unabhängig* vom Bewegungsverhalten der rufenden Teilnehmer sind.

- **Ruftyp III:** Rufe, die *unabhängig* vom aktuellen Aufenthaltsort des gerufenen Teilnehmers erfolgen und *abhängig* vom Bewegungsverhalten der rufenden Teilnehmer sind.
- **Ruftyp IV:** Rufe, die *abhängig* vom aktuellen Aufenthaltsort des gerufenen Teilnehmers erfolgen und *abhängig* vom Bewegungsverhalten der rufenden Teilnehmer sind.

Zu jedem der vier Ruftypen ($i \in \{I, II, III, IV\}$) definieren wir eine bedingte Rufratenmatrix ${}^{(i)}\mathbf{R}$ mit den Komponenten ${}^{(i)}r_{u,v}$. Aufgrund der Unabhängigkeit vom aktuellen Aufenthaltsort des gerufenen Teilnehmers bei den Ruftypen I und III gilt:

$${}^{(I)}r_{u,v} = {}^{(I)}r_{u,w} \quad \text{und} \quad {}^{(III)}r_{u,v} = {}^{(III)}r_{u,w} \quad \text{für alle } u, v, w \quad (5-52).$$

Zur Festlegung des Verhältnisses zwischen den Rufraten der verschiedenen Ruftypen definieren wir die nichtnegativen Ruftypenanteile ${}^{(i)}P_{\text{Ruf}}$, für die gelten muß

$${}^{(I)}P_{\text{Ruf}} + {}^{(II)}P_{\text{Ruf}} + {}^{(III)}P_{\text{Ruf}} + {}^{(IV)}P_{\text{Ruf}} = 1 \quad (5-53).$$

Darüberhinaus wird eine globale stationäre Rufrate Σ_r (bzw. $\Sigma_r^{(k)}$) zu einem Teilnehmer vorgegeben, die sicherstellen soll, daß

$$\sum_u \sum_v r_{u,v} = \sum_v r_v = \Sigma_r \quad (5-54).$$

Die Gesamtmatrix \mathbf{R} (bzw. $\mathbf{R}^{(k)}$) der bedingten Rufraten ergeben sich aus der Überlagerung der bedingten Rufratenmatrizen der vier Ruftypen:

$$\mathbf{R} = \Sigma_r \cdot \left({}^{(I)}P_{\text{Ruf}} \cdot {}^{(I)}\mathbf{R} + {}^{(II)}P_{\text{Ruf}} \cdot {}^{(II)}\mathbf{R} + {}^{(III)}P_{\text{Ruf}} \cdot {}^{(III)}\mathbf{R} + {}^{(IV)}P_{\text{Ruf}} \cdot {}^{(IV)}\mathbf{R} \right) \quad (5-55).$$

Die ${}^{(i)}r_{u,v}$ müssen also so gewählt werden, daß folgende Verhältnisse erfüllt werden

$$\begin{aligned} {}^{(I)}P_{\text{Ruf}} \cdot \sum_u \sum_v {}^{(I)}r_{u,v} &: {}^{(II)}P_{\text{Ruf}} \cdot \sum_u \sum_v {}^{(II)}r_{u,v} : \dots : {}^{(IV)}P_{\text{Ruf}} \cdot \sum_u \sum_v {}^{(IV)}r_{u,v} \\ &= {}^{(I)}P_{\text{Ruf}} : {}^{(II)}P_{\text{Ruf}} : {}^{(III)}P_{\text{Ruf}} : {}^{(IV)}P_{\text{Ruf}} \end{aligned} \quad (5-56),$$

woraus mit (5-54) folgt, daß

$$\sum_u \sum_v {}^{(i)}r_{u,v} = 1 \quad (5-57)$$

gelten muß.

Zur Berücksichtigung von „Distanzen“ definieren wir eine symmetrische Distanzmatrix Δ :

$$\Delta = \begin{bmatrix} d_{u,v} \end{bmatrix} \quad (5-58),$$

deren nichtnegativen Elemente $d_{u,v}$ einem Paar von Zuständen einen Wert, der als Distanz zwischen den Zuständen interpretiert wird, zuordnet. Dabei soll gelten:

$$d_{u,v} = d_{v,u} \text{ und } d_{u,u} = 0 \text{ für alle } u, v \quad (5-59).$$

Im weiteren soll die Matrix Δ als unabhängig von der Teilnehmerklasse k angenommen werden. Genausogut wäre jedoch auch denkbar, Distanzen zwischen Zuständen des Bewegungsmodells abhängig von der Teilnehmerklasse k zu definieren. Es ergäben sich dann die Matrizen $\Delta^{(k)}$ mit den Elementen $d_{u,v}^{(k)}$.

Um eine einheitliche Gewichtung der Distanzen vornehmen zu können, soll ferner gelten, daß die mittlere Distanz zwischen verschiedenen Zuständen den Wert 1 hat, dazu muß

$$\sum_{u \in S} \sum_{v \in S} d_{u,v} = N^2 - N \quad (5-60)$$

erfüllt sein.¹⁵ Um den Einfluß unterschiedlicher Distanzen auf die Rufraten festlegen zu können, wird ein *Spreizfaktor* δ definiert, für den $\delta \in [0, 1]$ gilt, und der die Abbildung von Distanzen auf Gewichtungsfaktoren steuert. Als Gewichtungsfunktion zur Abbildung von Distanzen auf Gewichtungsfaktoren wird die Funktion

$$f_{\delta}(d) = e^{-\frac{1-\delta}{\delta} \cdot d} \quad (5-61)$$

vorgeschlagen, die für $\delta = 1$ alle Distanzen unabhängig von ihrem Wert auf 1 abbildet, während für $\delta = 0$ alle Distanzen bis auf die Distanz 0 auf den Wert 0 abgebildet werden. Abbildung 5-6 zeigt die Gewichtungsfunktion $f_{\delta}(d)$ für verschiedene Werte von δ im Bereich von 0 bis 2, also bis zur doppelten mittleren Distanz zwischen den Zuständen.

Für die Generierung der Matrix \mathbf{R} auf Basis der bisher eingeführten Parameter werden hier nun zwei zueinander ähnliche Verfahren vorgestellt. Das erste Verfahren stützt sich auf die Definition eines „Heimatzustandes“ ab. Die grundlegende Idee ist bereits in [Sch97a] präsentiert. Das zweite Verfahren verallgemeinert das Konzept eines „Heimatzustandes“. Jedem Zustand wird ein „Heimatsfaktor“ zugordnet, der angibt zu welchem Grad der Zustand als „Heimatzustand“ zu betrachten ist.

15. N bezeichnet dabei die Gesamtzahl der Zustände.

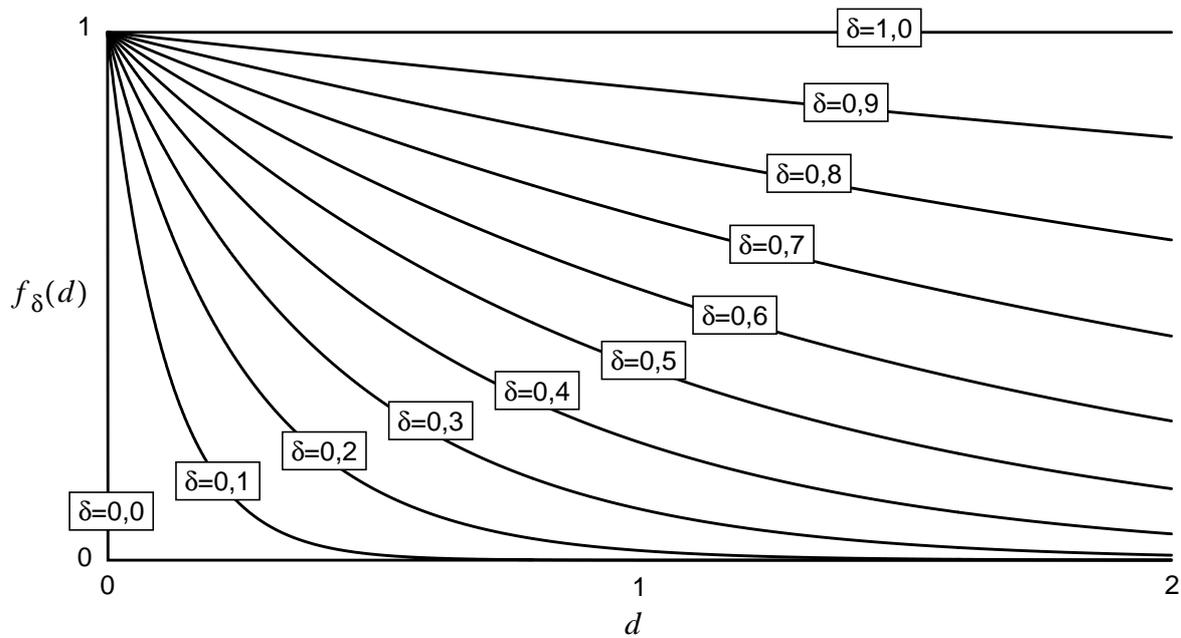


Abbildung 5-6: Gewichtungsfunktion $f_{\delta}(d)$ für verschiedene Werte des Spreizfaktors δ für Distanzen d zwischen 0 und 2 (= doppelte mittlere Distanz).

5.2.2.1 Verfahren mit Definition eines *Heimatzustandes*

Für jede Teilnehmerklasse k wird ein Heimatzustand $h^{(k)}$ definiert. Für die Rufraten werden je nach Ruftyp folgende Vereinbarungen getroffen (die in Klammern umgangssprachlich interpretiert werden):

- **Ruftyp I:** ${}^{(I)}r_{u,v}^{(k)}$ hängt vom Abstand zwischen u und $h^{(k)}$ ab. (Dies sind Rufe aus der Nähe der Heimat des Gerufenen).
- **Ruftyp II:** ${}^{(II)}r_{u,v}^{(k)}$ hängt vom Abstand zwischen u und v ab. (Dies sind Rufe aus der Nähe des aktuellen Aufenthaltsorts des Gerufenen).
- **Ruftyp III:** ${}^{(III)}r_{u,v}^{(k)}$ hängt davon ab, wieviele Teilnehmer der Klasse j sich im Mittel in u aufhalten und wie groß der Abstand zwischen $h^{(j)}$ und $h^{(k)}$ ist. (Dies sind Rufe von Teilnehmern, deren Heimat sich in der Nähe der Heimat des Gerufenen befindet).
- **Ruftyp IV:** ${}^{(IV)}r_{u,v}^{(k)}$ hängt davon ab, wieviele Teilnehmer der Klasse j sich im Mittel in u aufhalten und wie groß der Abstand zwischen $h^{(j)}$ und v ist. (Dies sind Rufe von Teilnehmern, deren Heimat sich in der Nähe des aktuellen Aufenthaltsorts des Gerufenen befindet).

Folgende Rufraten, die mit Hilfe der Distanzgewichtungsfunktion ermittelt werden, erfüllen die obigen Vereinbarungen:

$${}^{(I)}r_{u,v}^{(k)} = \frac{1}{{}^{(I)}C_v^{(k)}} \cdot f_{\delta}^{(k)}(d_{u,h^{(k)}}) \quad (5-62),$$

$${}^{(II)}r_{u,v}^{(k)} = \frac{1}{{}^{(II)}C_v^{(k)}} \cdot f_{\delta}^{(k)}(d_{u,v}) \quad (5-63),$$

$${}^{(III)}r_{u,v}^{(k)} = \frac{1}{{}^{(III)}C_v^{(k)}} \cdot \sum_{j \in K} \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{h^{(j)}, h^{(k)}}) \quad (5-64),$$

$${}^{(IV)}r_{u,v}^{(k)} = \frac{1}{{}^{(IV)}C_v^{(k)}} \cdot \sum_{j \in K} \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{h^{(j)}, v}) \quad (5-65),$$

wobei die Normierungsfaktoren ${}^{(i)}C_v^{(k)}$ sicherstellen sollen, daß die Summe der Rufraten zu einem Teilnehmer im Zustand v für alle v identisch gleich $1/N$ ist. Es muß also gelten:

$$\sum_{u \in S} {}^{(i)}r_{u,v}^{(k)} = \frac{1}{N} \quad (5-66),$$

was erfüllt wird durch:

$${}^{(I)}C_v^{(k)} = N \cdot \sum_{u \in S} f_{\delta}^{(k)}(d_{u, h^{(k)}}) \quad (5-67),$$

$${}^{(II)}C_v^{(k)} = N \cdot \sum_{u \in S} f_{\delta}^{(k)}(d_{u, v}) \quad (5-68),$$

$${}^{(III)}C_v^{(k)} = N \cdot \sum_{u \in S} \sum_{j \in K} \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{h^{(j)}, h^{(k)}}) \quad (5-69),$$

$${}^{(IV)}C_v^{(k)} = N \cdot \sum_{u \in S} \sum_{j \in K} \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{h^{(j)}, v}) \quad (5-70).$$

Bei den Ruftypen I, III und IV hängen die Rufraten entscheidend von der Definition des Heimatzustandes ab. Im folgenden Abschnitt wird das Konzept des „Heimatzustandes“ verallgemeinert.

5.2.2.2 Verfahren mit Definition von *Heimatfaktoren*

Anstatt einen einzelnen Zustand als Heimatzustand zu betrachten, wird nun eine Gruppe von Zuständen als „Heimatzustände“ definiert, und jedem dieser Zustände ein Heimatfaktor $\eta_u^{(k)}$ zugeordnet, der angibt zu welchem Grad ein Zustand „Heimatzustand“ ist. Für die Rufraten und die Normierungsfaktoren muß dann über alle Heimatzustände aufsummiert und entsprechend der Heimatfaktoren gewichtet werde. Es ergeben sich folgende Werte:

$${}^{(I)}r_{u,v}^{(k)} = \frac{1}{{}^{(I)}C_v^{(k)}} \cdot \sum_{w \in S} \eta_w^{(k)} \cdot f_{\delta}^{(k)}(d_{u, w})$$

mit

$${}^{(I)}C_v^{(k)} = N \cdot \sum_{u \in S} \sum_{w \in S} \eta_w^{(k)} \cdot f_{\delta}^{(k)}(d_{u, w}) \quad (5-71),$$

${}^{(II)}r_{u,v}^{(k)}$ wie in Gleichung (5-63) mit ${}^{(II)}C_v^{(k)}$ wie in Gleichung (5-68),

$${}^{(III)}r_{u,v}^{(k)} = \frac{1}{{}^{(III)}C_v^{(k)}} \cdot \sum_{w \in S} \sum_{j \in K} \sum_{z \in S} \eta_w^{(k)} \cdot \eta_z^{(j)} \cdot \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{z,w})$$

mit

(5-72),

$${}^{(III)}C_v^{(k)} = N \cdot \sum_{u \in S} \sum_{w \in S} \sum_{j \in K} \sum_{z \in S} \eta_w^{(k)} \cdot \eta_z^{(j)} \cdot \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{z,w})$$

$${}^{(IV)}r_{u,v}^{(k)} = \frac{1}{{}^{(IV)}C_v^{(k)}} \cdot \sum_{j \in K} \sum_{z \in S} \eta_z^{(j)} \cdot \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{z,v})$$

mit

(5-73).

$${}^{(IV)}C_v^{(k)} = N \cdot \sum_{u \in S} \sum_{j \in K} \sum_{z \in S} \eta_z^{(j)} \cdot \Pi_u^{(j)} \cdot f_{\delta}^{(k)}(d_{z,v})$$

Der Einfachheit halber wird im weiteren für alle Zustände häufig der jeweilige Heimatfaktor mit der zugehörigen Zustandswahrscheinlichkeit gleichgesetzt, also häufig $\eta_u^{(k)} = \pi_u^{(k)}$ gesetzt.

5.2.3 Zusammenfassung der Vorgehensweise

Abbildung 5-7 stellt zusammenfassend die Vorgehensweise zur Parametrisierung des Rufmodells dar. Die bedingte Rufratenmatrix $\mathbf{R}^{(k)}$ und die Verkehrsmatrix $\Lambda^{(k)}$ des Rufverkehrs werden mit Hilfe der Ruftypanteile ${}^{(i)}P_{\text{Ruf}}$, der globalen stationären Rufrate Σr und den bedingten Rufratenmatrizen ${}^{(i)}\mathbf{R}^{(k)}$ der vier Ruftypen ermittelt. Vorgaben für die Erstellung der ${}^{(i)}\mathbf{R}^{(k)}$ sind – neben der mittleren Zahl von Teilnehmern aller Teilnehmerklassen in allen Zuständen – die Distanzmatrix Δ , der Spreizfaktor $\delta^{(k)}$ und die Festlegung von Heimatzuständen ($h^{(j)}$) bzw. von Heimatfaktoren ($\eta_u^{(j)}$) für alle Teilnehmerklassen.

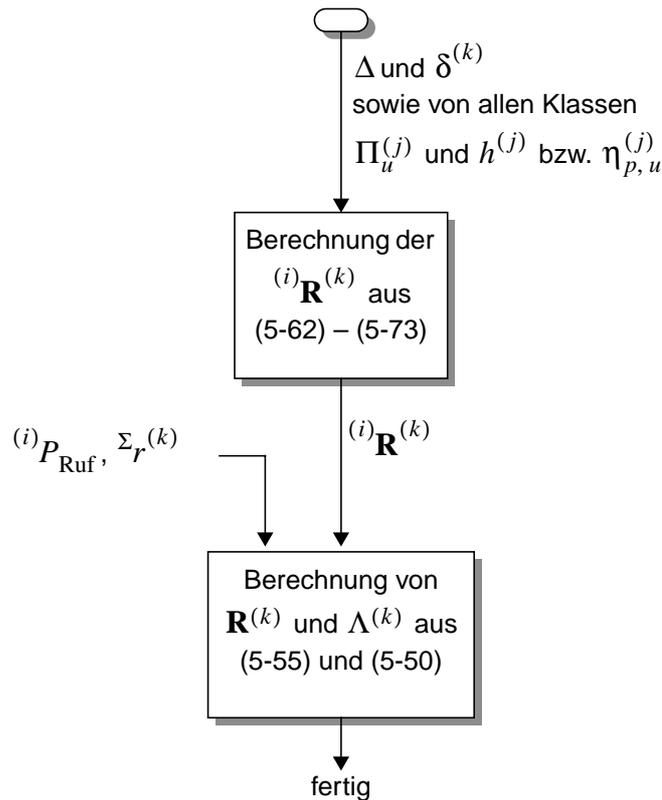


Abbildung 5-7: Vorgehensweisen zur Parametrisierung des Rufmodells für einen Teilnehmer der Klasse k unter Berücksichtigung aller Teilnehmerklassen j ($j \in K$).

Abbildung 5-8 zeigt ein Beispielszenario mit 625 Zuständen und 625 gleichgroßen Teilnehmerklassen, bei dem die Aufenthaltsorte in einem quadratischen Raster angeordnet sind und die Distanzen zwischen den Zuständen Manhattan-Distanzen sind. Jeder Teilnehmerklasse ist genau ein Aufenthaltsort als Heimatzustand zugeordnet (d.h. $h^{(k)} = k$). Die Zustandswahrscheinlichkeit eines Zustands hängt über $f_\delta(d)$ mit $\delta = 0,1$ vom Abstand zum Heimatzustand ab. Für die Teilnehmerklassen k und j sind die mittlere Zahl von Teilnehmern je Zustand ($\underline{\Pi}^{(k)}$ und $\underline{\Pi}^{(j)}$) in (a) und (b) abgebildet. Für Teilnehmer der Klasse k im Zustand v sind je Ruftyp in (c) bis (f) die entsprechenden Rufraten ${}^{(i)}r_v^{(k)}$ dargestellt. $\delta^{(k)}$ wurde dabei zu 0,3 gewählt. In (g) werden die Rufraten für Teilnehmer der Klasse k im Zustand v ($r_v^{(k)}$) bei gleichgewichtiger Überlagerung aller Ruftypen (${}^{(I)}P_{\text{Ruf}} = {}^{(II)}P_{\text{Ruf}} = {}^{(III)}P_{\text{Ruf}} = {}^{(IV)}P_{\text{Ruf}}$) gezeigt.

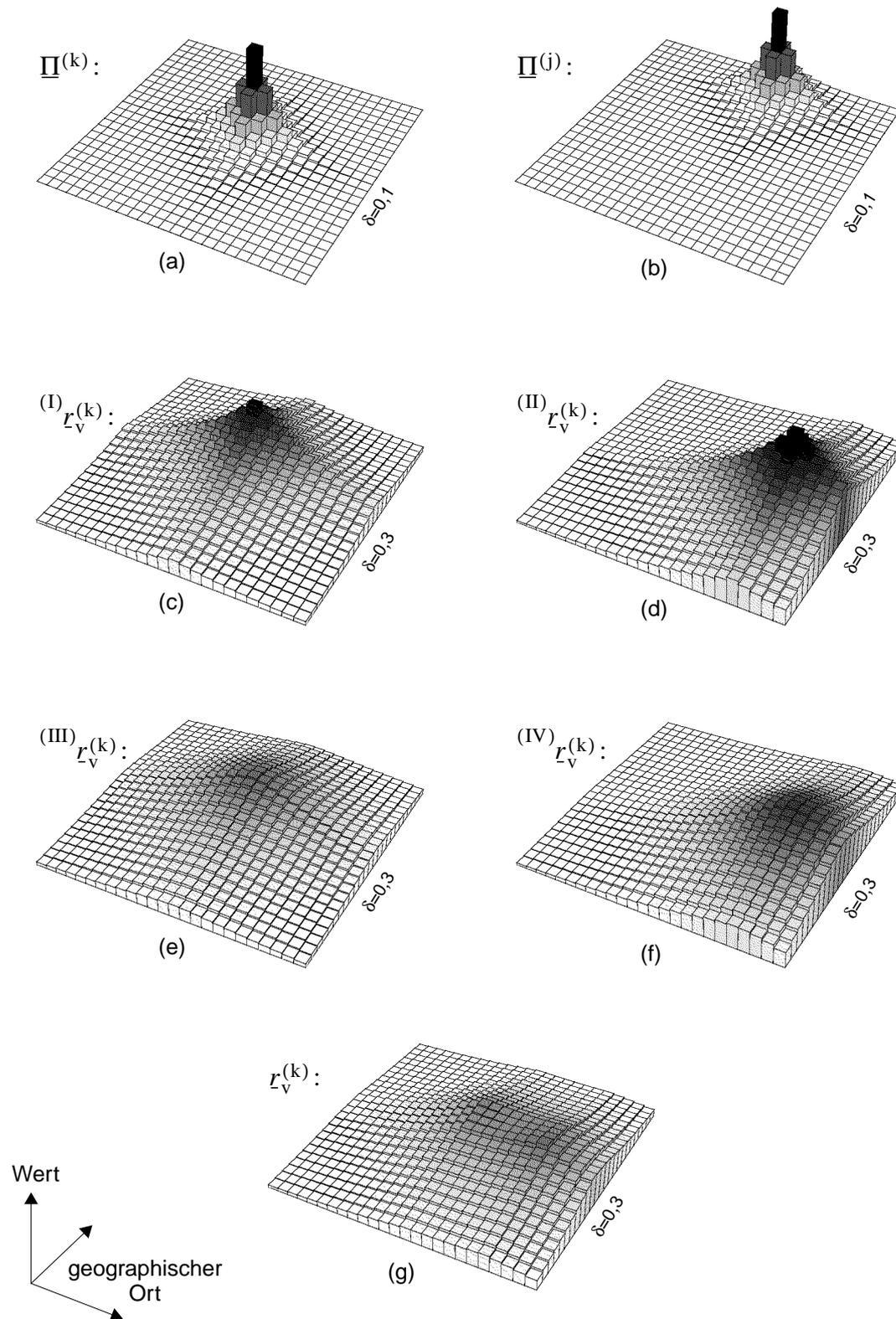


Abbildung 5-8: Beispielszenario 1: (a) mittlere Zahl von Teilnehmern der Klasse k je Zustand (b) mittlere Zahl von Teilnehmern einer beliebigen Klasse j je Zustand sowie (c) bis (f) Rufraten zu Teilnehmern der Klasse k im Zustand v für die Ruftypen I, II, III, IV und (g) Rufraten bei gleichgewichtiger Überlagerung aller Ruftypen.

Abbildung 5-9 zeigt das gleiche Szenario wie Abbildung 5-8, jedoch mit verändertem Bewegungsverhalten der Teilnehmer aller Klassen. Die Zustandswahrscheinlichkeit eines Zustands hängt hier über $f_\delta(d)$ mit $\delta = 0,3$ vom Abstand zum Heimatzustand ab. Die Rufraten für die Ruftypen III und IV werden dabei gleichmäßiger über alle Zustände verteilt. Die Rufraten für die Ruftypen I und II sind dagegen unabhängig vom Bewegungsverhalten der Teilnehmer.

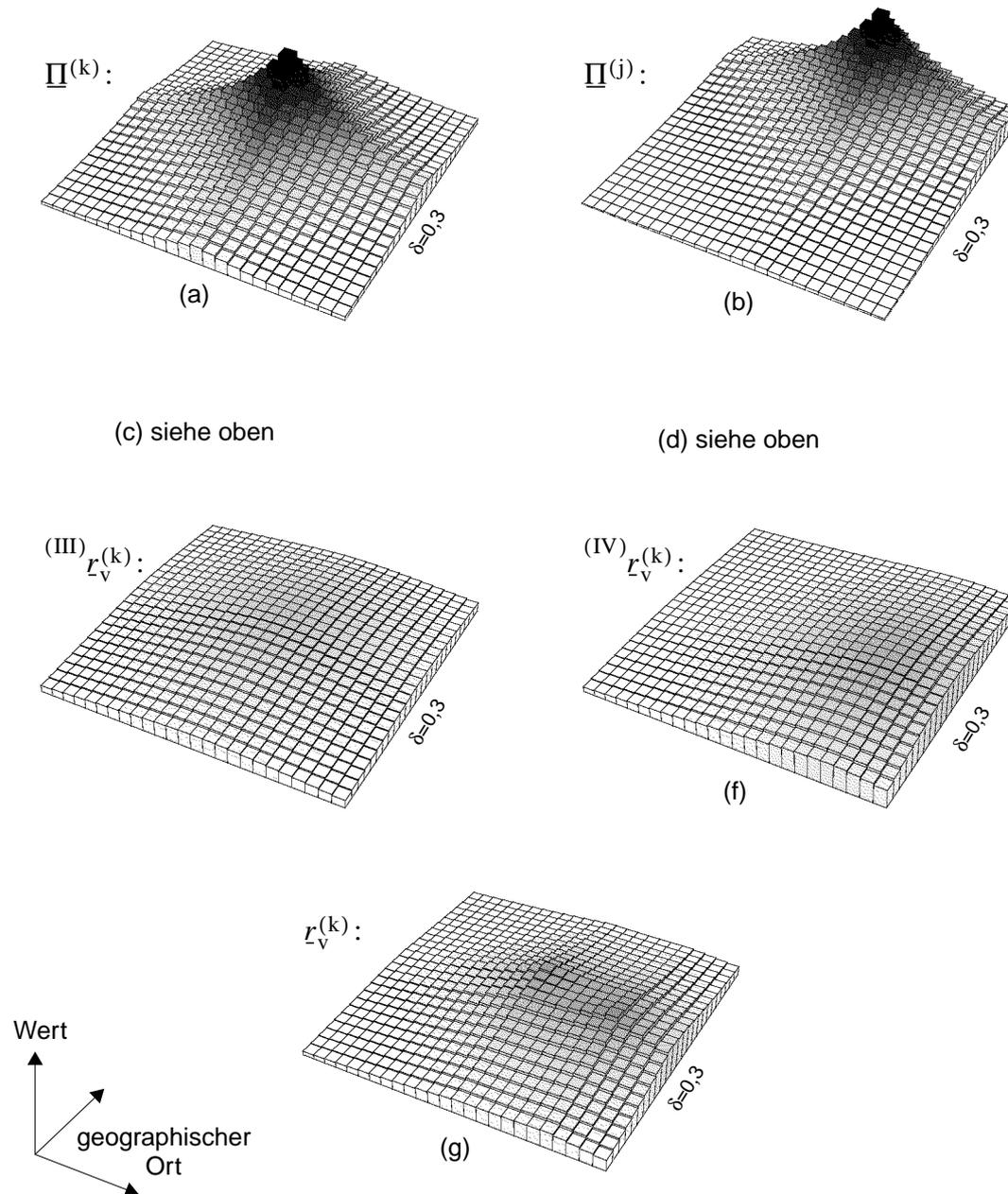


Abbildung 5-9: Beispielszenario 2: wie Beispielszenario 1, jedoch mit anderem Bewegungsverhalten der Teilnehmer. Teilbilder (c) und (d) sind nicht dargestellt, da sie identisch mit den entsprechenden Teilbildern in Abbildung 5-8 sind (Beschreibung der einzelnen Teilbilder siehe Abbildung 5-8).

5.3 Das Lastmodell

Mobilitätsverwaltung ist Teil der Dienstlogik, die in einem Kommunikationsnetz zur Verfügung gestellt wird. Je nach Art der Inanspruchnahme dieser Dienstlogik kommt es zu unterschiedlichen Protokollabläufen, welche die Verarbeitungs- und Kommunikationsinstanzen der Netzsteuerung mit unterschiedlicher Last belegen. Die Modellierung dieser Last kann – unabhängig von der Netztechnik und der tatsächlichen Realisierung – im Rahmen des konzeptionellen Modells des IN (INCM) vorgenommen werden. Erst auf der physikalischen Ebene des INCM müssen Unterscheidungen zwischen Netztechniken, wie zum Beispiel zwischen SS7-Signalisierung oder IP-basiertem Meldungs austausch, erfolgen.

5.3.1 Prinzipielle Vorgehensweise

In [Baf95, BS95, BS96] wird eine Vorgehensweise zur Ermittlung der Last, die durch Dienstlogik in „Intelligenten Netzen“ erzeugt wird, präsentiert. Dabei werden speziell für das IN CS-1 verschiedene Dienstlogikszenerarien in der globalen funktionalen Ebene durch Verkettung sogenannter „Sub-SIBs“ dargestellt. Sub-SIBs stellen eine Verfeinerung von SIBs dar, resultieren aus der Kombination aller möglichen logischen Abläufe eines SIB und sind direkt auf Vorgänge in der verteilten funktionalen Ebene abbildbar. Die Abbildung auf Aktionen (FEAs) und Informationsfluß in der verteilten funktionalen Ebene wird dabei unabhängig von der konkreten Netzarchitektur, d. h. von der Zuordnung von funktionalen Instanzen zu physikalischen Instanzen in der physikalischen Ebene des INCM, gehalten. Welche Folge von Sub-SIBs einem Dienstlogikszenerario zuzuordnen ist, hängt im wesentlichen von den durch den Dienstleister vorgegebenen SIB-Parametern und dem Verhalten der Teilnehmer bei Benutzung des Dienstes ab, ist also meist ebenfalls unabhängig von der Netzarchitektur. Diese wird erst bei der Abbildung auf Vorgänge in der physikalischen Ebene relevant, wobei gegebenenfalls auch Informationsfluß und Aktionen ignoriert werden können, z. B. wenn aus ihnen aufgrund der Platzierung von funktionalen Instanzen am selben Ort keine Meldungen oder Bearbeitungsprozesse resultieren.

Sub-SIBs unabhängig von der Netzarchitektur zu definieren, ist jedoch nicht immer ganz einfach. Der SIB „*User Interaction*“ im IN CS-1 ist dafür ein gutes Beispiel. Hier hängen die Vorgänge in der verteilten funktionalen Ebene von der Verteilung der funktionalen Instanzen in der physikalischen Ebene ab. Dies rührt daher, daß SRF-Funktionalität in einem SSP integriert oder nur über einen bestimmten SSP zugänglich sein kann. Um die SRF-Funktionalität in diesen Fällen auch für Rufe an anderen SSPs verfügbar zu machen, sieht [Q.1214] eine „*service assist procedure*“ und eine „*service hand-off procedure*“ vor, bei denen die Rufsteuerung zeitweise oder gänzlich an eine SSF im unterstützenden SSP abgegeben wird. In [Kan94] werden deshalb spezielle Sub-SIBs für solche Fälle vorgesehen. Damit wird jedoch die Auswahl der Sub-SIBs zur Beschreibung eines Dienstlogikszenerarios von der spezifischen Netzarchitektur abhängig.

Die Lastmodellierung nach [Baf95] ist prinzipiell auch auf die verteilte Mobilitätsverwaltung anwendbar. Die konkrete Ausgestaltung der Vorgehensweise, die auf das IN CS-1 abgestimmt ist und in [Kan94] dokumentiert ist, bedarf jedoch der Erweiterung. Hierfür gibt es im wesentlichen vier Ursachen:

- ❑ Bei einigen Verfahren der Mobilitätsverwaltung kommt es zu einer ausgeprägten Parallelisierung von Vorgängen. Daher können die Abläufe in der funktionalen Ebene nicht mehr als serielle Abfolge modelliert werden.
- ❑ Bei einigen Verfahren der Mobilitätsverwaltung sind die Vorgänge in der funktionalen Ebene stark von der Datenbankstruktur und der aktuellen Platzierung von Daten abhängig. Eine einfache netzarchitekturunabhängige Definition von Sub-SIB wird dadurch deutlich schwieriger.
- ❑ Die Abbildung auf die physikalische Ebene ist für Mobilitätsverwaltung komplexer und von einer größeren Zahl von Parametern abhängig. In [Kan94] ist lediglich vorgesehen, jedem SSP je Dienstszenario mit prozentualen Anteilen gewichtete SCP-IP-SDP-Tupel zuzuordnen.
- ❑ Unabhängig von der Modellierung von Mobilitätsverwaltung muß das Verfahren nach [Kan94] erweitert werden, wenn damit Abläufe im IN CS-2 modelliert werden sollen. So ist in [Kan94] beispielsweise Informationsfluß zwischen SCFs oder zwischen SDFs nicht vorgesehen.

Die Verfahrensweise nach [Kan94] ist zwar zum Teil in der Lage, die oben genannten Charakteristika von Mobilitätsverwaltungsszenarien näherungsweise nachzubilden, führt jedoch zu einem sehr unübersichtlichen und komplexen Vorgehen. So ist zum Beispiel vorgesehen, Teilszenarien des Dienstlogikablaufs mit jeweils eigenen SCP-IP-SDP-Tupeln in verschiedenen Folgen, sogenannten „Rufszszenarien“, kombinieren zu können. Man könnte zwar Informationsfluß zwischen einer SCF und mehreren verschiedenen SDFs je nach beteiligten SDPs als unterschiedliche Rufszszenarien definieren, je nach Zahl der SDPs würde man aber sehr schnell eine sehr große Zahl von Kombinationen und damit von Rufszszenarien erhalten. Ebenso könnte man parallele Vorgänge als unabhängige Rufszszenarien betrachten und die Leistungsbewertung anhand der Ergebnisse der unabhängigen Betrachtung vornehmen, was zwar das Lastmodell verfälschen würde, unter Umständen jedoch näherungsweise gültige Ergebnisse bei der Leistungsbewertung liefern dürfte.

5.3.2 Angepaßte Vorgehensweise

Hier soll nun eine modifizierte Modellierung präsentiert werden, die zum einen speziell auf die Spezifika der Mobilitätsverwaltung abgestimmt ist und zum anderen eine striktere Trennung zwischen den verschiedenen Ebenen des konzeptionellen Modells des IN vornimmt. In der globalen funktionalen Ebene wird die Dienstlogik möglichst unabhängig von den Verfahren zur Mobilitätsverwaltung bzw. zur Datenhaltung beschrieben. In der verteilten funktionalen Ebene wird eine ausreichend detaillierte Beschreibung angestrebt, die zwar abhängig vom verwendeten Verfahren zur Mobilitätsverwaltung ist, jedoch unabhängig von der konkreten Netz-

architektur, also beispielsweise auch unabhängig von der Zahl der Hierarchiestufen und der Knoten in einem verteilten Datenbanksystem. In der physikalischen Ebene schließlich müssen die konkreten Meldungen und Aktionen, wie sie durch die Netzarchitektur und durch die zugrundeliegende Netztechnik bestimmt sind, dargestellt werden.

Auf der *globalen funktionalen Ebene* werden, wie im IN CS-1, einfache SIBs verwendet, die beispielsweise Daten lesen, schreiben oder verändern. Einzelne Dienstszenarien werden weiterhin als Folge von Sub-SIBs dargestellt. Die Abbildung von Sub-SIBs auf Abläufe in der verteilten funktionalen Ebene erfolgt jedoch nicht mehr direkt aus der Definition der Sub-SIBs, sondern berücksichtigt die verwendeten Verfahren zur Mobilitätsverwaltung bzw. zur Datenerhaltung.

Die Abfolge von Aktionen und der Austausch von Informationsflußkomponenten in der *verteilten funktionalen Ebene* soll hier durch erweiterte *Message Sequenz Charts* (MSC) nach [Z.120] dargestellt werden. Diese werden im nächsten Unterkapitel (5.3.3) vorgestellt und um zusätzliche Beschreibungsmittel ergänzt. Entsprechend erweitert ermöglichen sie eine Lastbeschreibung, die unabhängig ist von der konkreten Netzarchitektur, also auch unabhängig von der Zahl der Hierarchiestufen und der Knoten in einem verteilten Datenbanksystem.

Die Abbildung der Lastbeschreibung auf die *physikalische Ebene* schließlich berücksichtigt die tatsächliche Netzarchitektur, die tatsächliche Datenverteilung, konkrete Signalisierungsprotokolle und Wegführung. Hier werden auch zusätzliche Protokollfunktionen berücksichtigt, die zur Kommunikation zwischen Netzknoten (Signalisierung) notwendig sind. Für auszutauschende Meldungen wird die Meldungslänge bestimmt. Für Last, die durch Prozessoren bearbeitet wird, werden quantitative Angaben gemacht, die es erlauben, die Bearbeitungszeit in einem bearbeitenden Prozessor zu bestimmen. Für Bearbeitungszeiten in Prozessoren müssen geeignete Messungen oder Schätzungen vorliegen. In der Regel kann man bei den beanspruchten Belegungsdauern von deterministischen Werten ausgehen. Einige Bearbeitungsdauern, wie zum Beispiel die von Suchvorgängen, können jedoch besser durch (unabhängige) Zufallsvariablen beschrieben werden. Diese werden dann durch Verteilungsfunktionen und/oder durch Momente (Mittelwert und höhere Momente) charakterisiert.

Für jedes Szenario der verteilten funktionalen Ebene ergeben sich eine Vielzahl von möglichen Szenarien in der physikalischen Ebene.¹⁶ Für jedes Szenario erhält man eine Lastbeschreibung, die eine Folge von Ressourcenanforderungen darstellt, bei der es auch zu einer Aufspaltung in mehrere parallel verlaufende Zweige kommen kann. In Anlehnung an den Begriff *functional entity action* wird solch eine Ressourcenanforderung hier als *physical entity action* (PEA) bezeichnet. Eine PEA wird durch die Art der Last (Meldung oder Prozessorbelegung), Angaben über den Ort der Belastung und quantitative Angaben über die Größe der Last näher

16. Es ist daher sinnvoll, diese Abbildung im Rahmen eines Software-Werkzeugs durchzuführen und zu verwalten.

beschrieben. Abbildung 5-10 zeigt das Beispiel eines Lastszenarios. Nach der PEA₈ ist ein Abschlußsymbol angefügt, das für die Leistungsanalyse anzeigt, daß nach Bearbeitung dieser PEA das Szenario als abgearbeitet betrachtet werden kann.

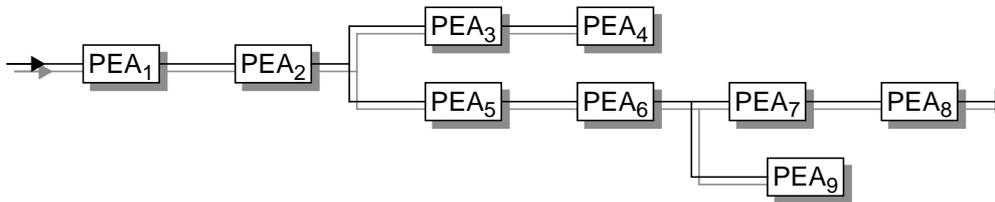


Abbildung 5-10: Beispiel eines Lastszenarios in Form einer Folge von Ressourcenbelegungen (PEAs), bei der es auch zu Aufspaltungen in parallel verlaufende Zweige kommen kann.

5.3.3 Zusätzliche Beschreibungsmittel

Für das IN CS-1 sind sehr einfache *Message Sequenz Charts* (MSC) ausreichend. Es genügen die Grundelemente *instance*, *message* und *action*.¹⁷ Für verteilte Mobilitätsverwaltung und Datenhaltung müssen dagegen mächtigere Beschreibungsmittel verwendet werden. Will man unabhängig von der Netzarchitektur solche Abläufe darstellen, so benötigt man Konzepte wie sie in [Z.120] beispielsweise in Form sogenannter *inline expressions* zur Verfügung stehen, welche unter anderem alternative (*alt*) und parallele (*par*) Abläufe sowie Iterationen in Form von Schleifen (*loop*) vorsehen. Damit kann zum Beispiel der SIB „*User Interaction*“ auf einen MSC mit alternativen Teilen bezüglich der *service assist procedure*, der *service hand-off procedure* und dem Ablauf bei direkt zugänglicher SRF, zwischen denen erst bei der Abbildung auf die physikalische Ebene ausgewählt wird, abgebildet werden. Weitere wünschenswerte Beschreibungsmittel fehlen jedoch noch im aktuellen Standard. So sind Schleifen zur Zeit nur in Form von sequentiellen Folgen definiert, und eine Meldung gleichzeitig an mehrere Instanzen zu schicken (*multicast*), ist nicht formalisiert. In [SG10Q9] werden jedoch Vorschläge zu entsprechenden Erweiterungen ansatzweise diskutiert. Abbildung 5-11 zeigt den Vorschlag zur graphischen Darstellung von *Multicast*-Meldungen, welcher im folgenden verwendet werden wird. Bisher ist auch noch keine variable Zahl von gleichartigen Instanzen vorgesehen. Dies ist

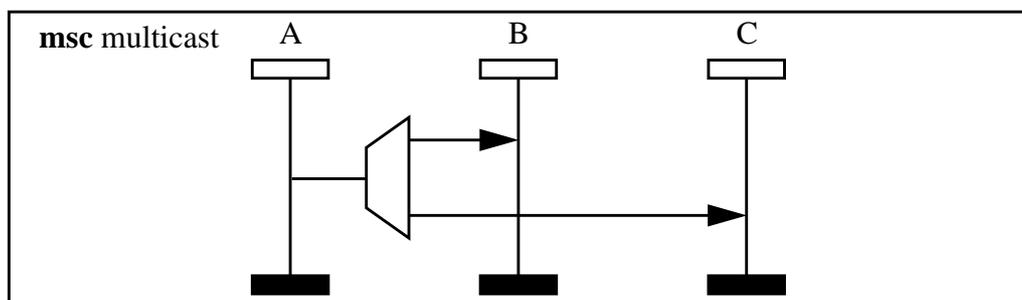


Abbildung 5-11: Mögliche Erweiterung von [Z.120] um *Multicast*-Meldungen nach [SG10Q9].

17. Wobei diese um Attribute bezüglich Länge oder Bearbeitungsaufwand erweitert sein sollten.

jedoch für die netzarchitekturunabhängige Beschreibung der Mobilitätsverwaltung von großer Wichtigkeit. Will man etwa den Ablauf bei Verwendung der in Unterkapitel 3.3.2 beschriebenen Techniken *Referral*, *Chaining*, *Multichaining* und *Passing* darstellen, so sollte dies unabhängig von der Zahl der weiterreichenden Instanzen geschehen können. Im Rahmen dieser Arbeit kann kein generelles Konzept zur Erweiterung von [Z.120] um eine variable Zahl von gleichartigen Instanzen erarbeitet werden. Für die speziellen Anforderungen der Netzarchitektur-unabhängigen Beschreibung von Vorgängen in der verteilten funktionalen Ebene bei der Verwendung unterschiedlicher Datenhaltungsstrategien wird jedoch eine Darstellungsform vorgeschlagen und präsentiert, die zumindest für die beschriebenen Weiterreichungstechniken geeignet ist. Es werden zwei wesentliche Erweiterungen eingeführt: Instanzen können einen kleingeschriebenen Index (z.B. „ n “) oder einen großgeschriebenen Index (z.B. „ N “) tragen, und Meldungen von oder zu diesen Instanzen können mit Adressen versehen werden. Instanzen mit kleinem Index stehen repräsentativ für eine variable Zahl von gleichartigen Instanzen, die sich im dargestellten MSC identisch verhalten. Eine Instanz mit großgeschriebenem Index und demselben Namen wie eine Instanz mit kleinem Index ist zwar von gleicher Art, verhält sich jedoch im dargestellten MSC anders. Die Grundannahme besteht darin, daß bei einem Ablauf mit einer variablen Zahl von gleichartigen Instanzen sich $N-1$ Instanzen identisch verhalten, während die N -te Instanz von deren Verhalten abweicht und dadurch das Einbeziehen weiterer gleichartiger Instanzen beendet. Der Fall $N=1$ ist ein Sonderfall bei dem Meldungen zu den (nicht-vorhandenen) klein indizierten Instanzen entweder ignoriert oder virtuell weitergereicht werden. Der Austausch von Meldungen zwischen den gleichartigen Instanzen wird durch die relative Adressierung in der Form $n+1$ oder $n-1$ dargestellt. Der gedachte Index n wird dabei entsprechend erhöht oder erniedrigt. Hat er weder 0 noch eine obere Grenze N erreicht, so geht die Meldung wieder an eine Instanz mit Index n , und derselbe Ablauf, der nach dem Empfang der letzten so adressierten Meldung abgelaufen ist, wiederholt sich. Ist dagegen 0 oder N erreicht, so geht die Meldung an die Instanz, für die sie nach der üblichen Konvention bestimmt ist. Meldungen zu allen gleichartigen Instanzen mit identischem Verhalten werden mit der absoluten Adresse wie beispielsweise $\{0..N\}$ versehen. Innerhalb eines *loop inline expression operators* nach [Z.120] bezeichnet die Adresse n die Instanz, die im momentanen Schleifendurchlauf adressiert wird. Eingerahmt werden all diese Abläufe durch ein *inline expression symbol* das mit dem neuen Operatorbezeichner „var“ versehen ist. Die Abbildungen 5-12 bis 5-15 zeigen die Verwendung der vorgeschlagenen Erweiterungen (schattiert dargestellt) zur Beschreibung der Weitergabetechniken *Verweisen*, *Verketteten*, *Abgeben* und *paralleles Multichaining*.¹⁸ Abbildung 5-16 zeigt wiederholtes Verwenden des vorgeschlagenen *inline expression operators*. Auf eine formale Spezifikation der grafischen oder einer textuellen Darstellung wird hier verzichtet, dies kann jedoch analog zur Spezifikation in [Z.120] erfolgen.

18. Aktionen, Meldungsaustausch mit der Umgebung und präzisierende Operanden sind nicht explizit dargestellt.

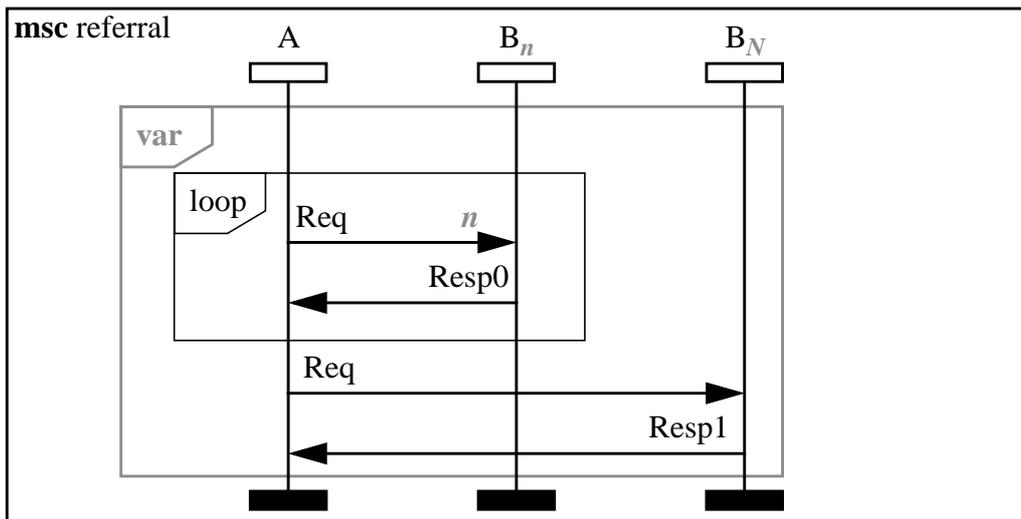


Abbildung 5-12: Erweiterter *Message Sequenz Chart* zur Beschreibung der Weitergabetechnik *Verweisen (Referral)*.*

* Die Weitergabetechnik „sequentielles *Multichaining*“ (*Polling*) würde genauso dargestellt werden.

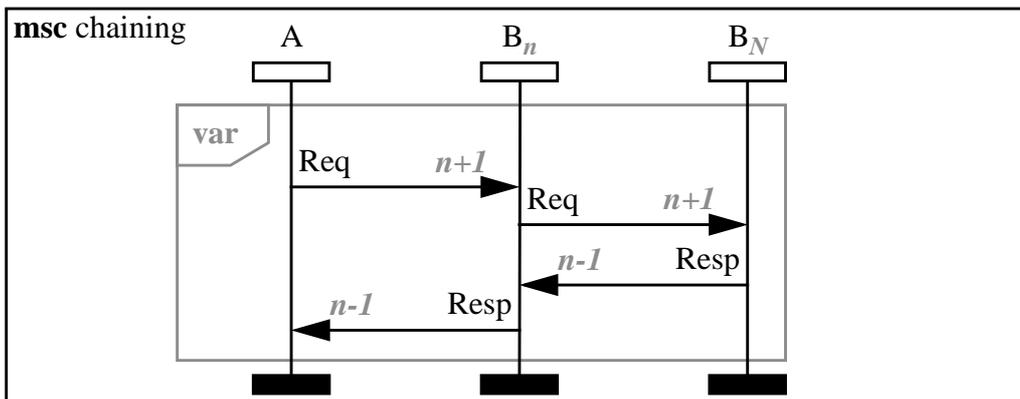


Abbildung 5-13: Erweiterter *Message Sequenz Chart* zur Beschreibung der Weitergabetechnik *Verketteten (Chaining)*.

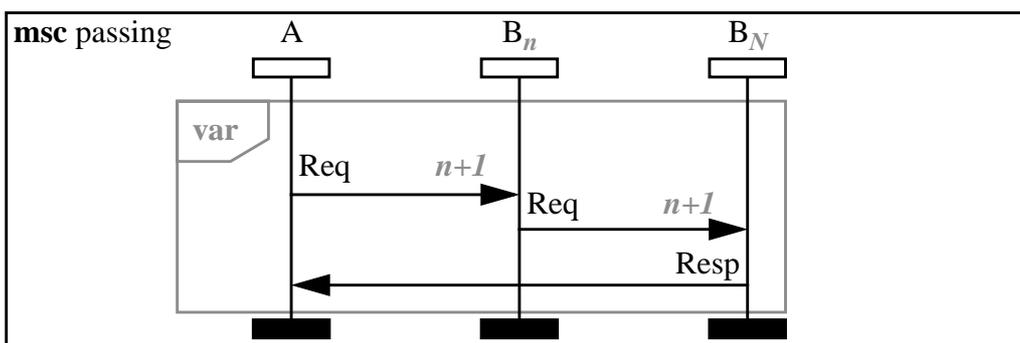


Abbildung 5-14: Erweiterter *Message Sequenz Chart* zur Beschreibung der Weitergabetechnik *Abgeben (Passing)*.

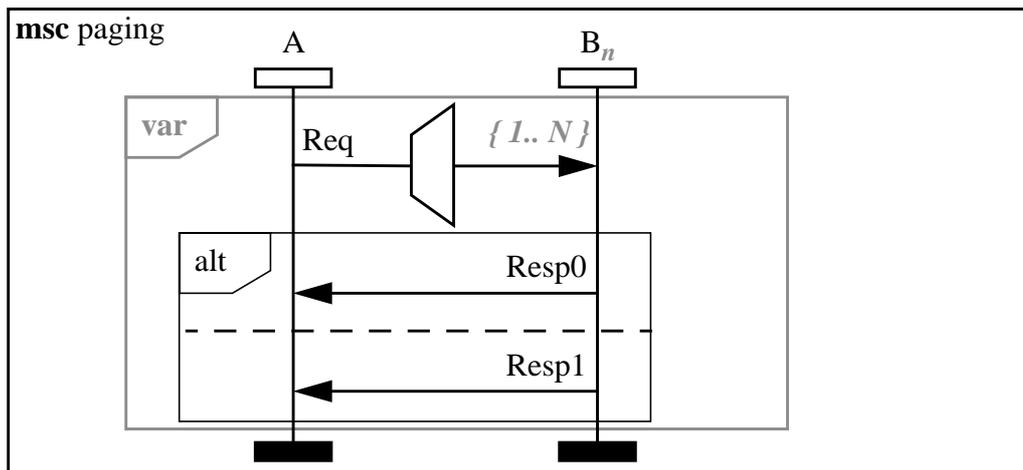


Abbildung 5-15: Erweiterter *Message Sequenz Chart* zur Beschreibung der Weitergabetechnik „paralleles Multichaining“ (*Paging*).

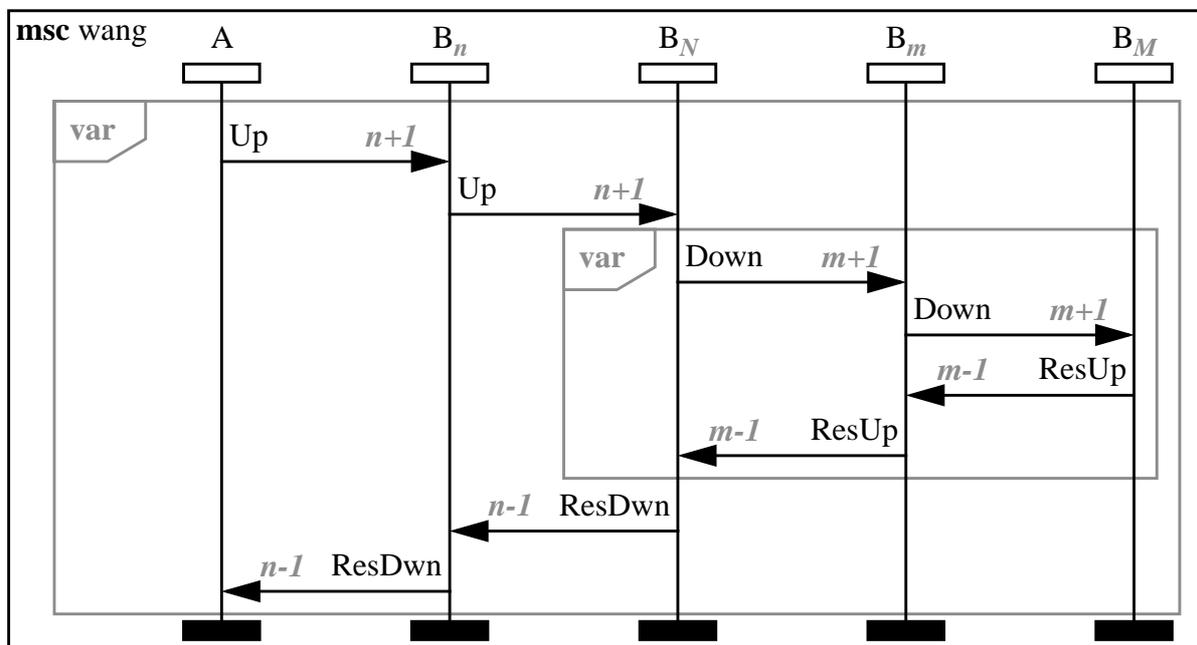


Abbildung 5-16: Erweiterter *Message Sequenz Chart* mit wiederholter Verwendung des vorgeschlagenen *inline expression operators* „var“.

Zur weiteren Erläuterung werden für $N=1$, $N=2$ und $N=3$ in Abbildung 5-17 die gewöhnlichen *Message Sequenz Charts* zur Beschreibung der Weitergabetechnik *Verketteten* dargestellt, die aus der Darstellung nach Abbildung 5-13 hergeleitet werden können.

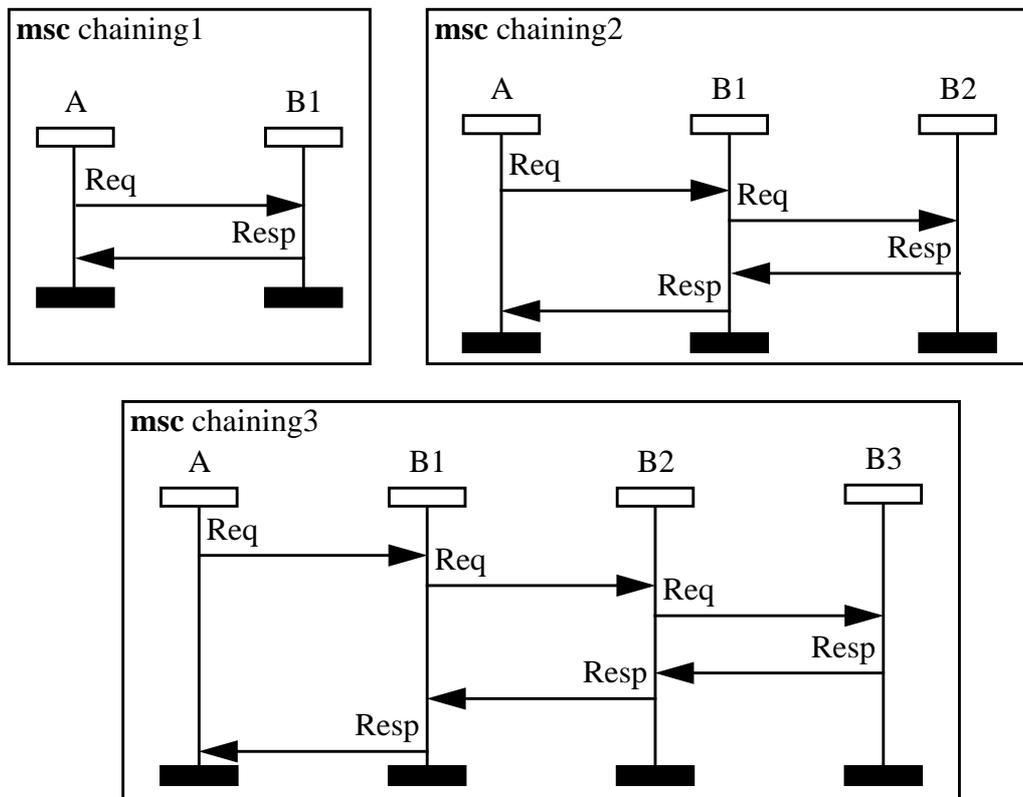


Abbildung 5-17: Gewöhnliche *Message Sequenz Charts* zur Beschreibung der Weitergabetechnik *Verketteten* (*Chaining*) für $N=1$, $N=2$ und $N=3$ beteiligte, gleichartige Instanzen.

5.4 Das Netzmodell

5.4.1 Grundlegende Modellkomponenten

Im Rahmen des oben beschriebenen Lastmodells wird jedes mögliche Szenario durch eine Folge von Ressourcenanforderungen bzw. Bearbeitungsschritten („*Physical Entity Actions*“) dargestellt. Als Ressourcen werden Prozessoren in Netzknoten und Übertragungstrecken betrachtet. Für jeden Bearbeitungsschritt läßt sich die beanspruchte Ressource und die Dauer der beanspruchten Belegung ermitteln. Diese Dauer ergibt sich bei Übertragungstrecken aus der Länge der zu übertragenden Meldung und der verfügbaren Übertragungsrates. Für Bearbeitungszeiten in Prozessoren sind die notwendigen Angaben Teil des Lastmodells und müssen gegebenenfalls noch an den speziellen Typ von Prozessor angepaßt werden.

Während das Lastmodell die Ressourcenanforderungen für jedes mögliche Szenario getrennt und unabhängig beschreibt, muß im Netzmodell die Überlagerung der Verkehre der verschiedenen Szenarien und die Behandlung derselben berücksichtigt werden. Neben dem

Nachrichtenverkehr, der aus dem Lastmodell ermittelt wird, kann auch zusätzlicher Hintergrundverkehr modelliert werden, der mit dem Verkehr aus dem Lastmodell um die vorhandenen Ressourcen konkurriert. Außer den für das Lastmodell relevanten Ressourcen müssen im Netzmodell weitere Ressourcen und Modellkomponenten berücksichtigt werden. Dies sind vor allem Ressourcen, die belegt werden, weil eine für das Lastmodell relevante Ressource nicht immer sofort bei Bedarf zur Verfügung steht, sondern erst dann, wenn sie einer Anforderung zugewiesen wird. Solch eine Ressource kann man als Puffer, Warteschlange oder Speicherplatz bezeichnen. Die Verfahren zur Zuweisung von Ressourcen zu Anforderungen („*Scheduling*“) sind Teil des Netzmodells. Diese Verfahren können zum Beispiel mit Hilfe von Prioritäten die Reihenfolge der Bearbeitung von Anforderungen festlegen und diese gegebenenfalls auch unterbrechen. Auch durch sie kann eine Netzressource zusätzlich belegt werden. So können die Unterbrechung der Bearbeitung einer Anforderung und die spätere Wiederaufnahme zusätzliche Bearbeitungszeit und Speicherplatz beanspruchen.

Die wesentlichen Modellkomponenten des Netzmodells sind:

- Bedieneinheiten
- Warteschlangen
- Verzögerungsglieder

Sie werden dazu verwendet, Netzknoten und Übertragungsabschnitte zu modellieren. In Abbildung 5-18 ist das Modell eines Übertragungsabschnitts dargestellt. Die Warteschlange dient zur Modellierung der Wartezeit, die vergeht, bis mit der Übertragung einer Meldung begonnen werden kann. Die Bedieneinheit modelliert die Übertragungszeit in Abhängigkeit von der Meldungslänge, und das Verzögerungsglied modelliert die Laufzeit auf dem Übertragungsabschnitt. Ein Netzknoten wird durch eine Reihe von Warteschlangen, Bedieneinheiten und Verzögerungsgliedern modelliert. Darüberhinaus muß auch das Verfahren zur Zuweisung von Ressourcen in das Modell eines Netzknotens miteinbezogen werden. Im folgenden Abschnitt soll diejenige Möglichkeit zur Berücksichtigung derartiger Verfahren vorgestellt werden, die im restlichen Teil der Arbeit verwendet werden wird.

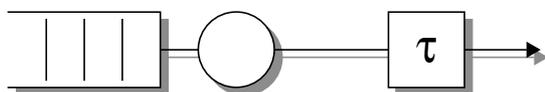


Abbildung 5-18: Modell eines Übertragungsabschnitts.

5.4.2 Modellierung der Netzknoten

Innerhalb eines Netzknotens steht im allgemeinen eine größere Zahl von Prozessoren zur Verfügung, um Anforderungen zu bearbeiten. Einige dieser Prozessoren führen spezielle Aufgaben durch. Beispielsweise kann es Prozessoren geben, die für eine Gruppe von Signaliserverbindungen die zugehörigen Protokollfunktionen des Signalisiersystems realisieren. Andere Prozessoren übernehmen ein breiteres Spektrum von Aufgaben. Häufig wird zwischen mehreren Prozessoren Lastteilung durchgeführt, so daß Anforderungen gleichmäßig auf mehrere

Prozessoren verteilt werden. Die Verfahren zur Verteilung der Anforderungen auf die verschiedenen Prozessoren sind ein wesentlicher Teil des Modells eines Netzknotens. Für eine simulative Leistungsbewertung lassen sich solche Verfahren relativ leicht algorithmisch nachbilden. Für eine analytische Leistungsbewertung müssen im allgemeinen gewisse Vereinfachungen vorgenommen werden. In dieser Arbeit soll ein Verfahren modelliert werden, das neben einer einfachen simulativen Nachbildung auch eine einfache analytische Behandlung erlaubt, dabei jedoch noch genügend Spielraum läßt, um das Verhalten tatsächlich realisierter Verfahren anzunähern. Die Verteilung von Anforderungen auf Prozessoren erfolgt dabei in zwei Schritten. Trifft eine Anforderung an einem Netzknoten ein oder wird nach Bearbeitungsende durch einen Prozessor Bearbeitungskapazität bei anderen Prozessoren nachgefragt, so wird eine Gruppe von Bearbeitungsschritten zusammengefaßt und zufällig ein Prozessor aus der Gruppe der zuständigen Prozessoren ausgewählt. Dieser Prozessor ist dann für die Bearbeitung der Gruppe von Bearbeitungsschritten zuständig. Im zweiten Schritt konkurrieren alle Anforderungen, die im ersten Schritt einem Prozessor zugewiesen werden, um die Bedienung. Dabei werden Prioritäten zwischen verschiedenen Bearbeitungsschritten berücksichtigt, die sich entweder als Warteprioritäten oder auch als Unterbrechungsprioritäten auswirken.

Abbildung 5-19 illustriert den ersten Schritt des betrachteten Ressourcenzuweisungsverfahrens. Eine Folge von Bearbeitungsschritten eines Lastszenarios wird in Gruppen zerteilt. Die Bearbeitung einer solchen Gruppe wird einem einzelnen Prozessor zugewiesen. Die Auswahl des zuständigen Prozessors erfolgt zufallsgesteuert. Bei jeder Auswahlmöglichkeit wird ausgewürfelt, welcher Prozessor gewählt werden soll. Die Auswahlwahrscheinlichkeit für einen Prozessor wird konstant vorgegeben.

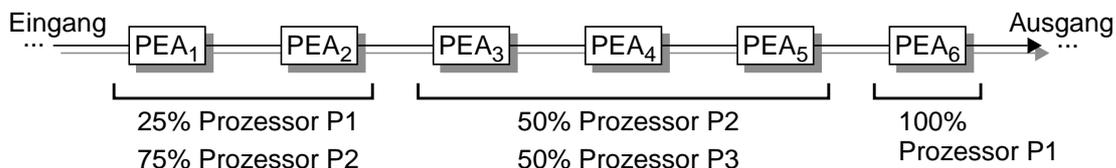


Abbildung 5-19: Beispiel einer Zuweisung von Gruppen von Bearbeitungsschritten (*Physical Entity Actions*, PEA) an Prozessoren.

Abbildung 5-20 illustriert die Prozessorenauswahl anhand von Verzweigungen, die mit Wahrscheinlichkeiten versehen sind. Am Eingang in den Netzknoten und nach jedem Prozessor sind solche Verzweigungen vorgesehen. Diese Darstellung ist prinzipiell mächtiger als die Darstellung in Abbildung 5-19, da sich die Zweige nach einer Auswahl unabhängig voneinander weiterentwickeln können. Anzumerken ist darüberhinaus, daß es zu einer Rückführung zum selben Prozessor (Prozessor P2) zwischen PEA₂ und PEA₃ kommen kann, was im analytischen Modell explizit berücksichtigt werden muß.

Obwohl die oben beschriebene Zuordnung von Bearbeitungsschritten zu Prozessoren relativ mächtig ist und weder bei der simulativen noch bei der analytischen Behandlung prinzipielle Schwierigkeiten bereitet, werden hier für die weitere Arbeit einige Einschränkungen gemacht. Das beschriebene Zuordnungsverfahren führt dazu, daß Bearbeitungsschrittfolgen eines Last-

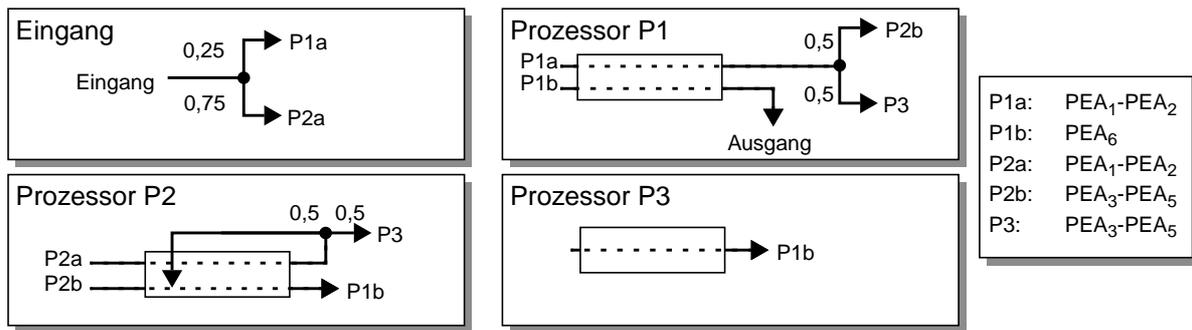


Abbildung 5-20: Darstellung der Zuweisung nach Abbildung 5-19 mit Hilfe von Verzweigungen (\rightarrow), die mit Wahrscheinlichkeiten versehen sind.

szenarios je nach Zuordnungsvariante sehr unterschiedliche mittlere Durchlaufzeiten erfahren. Dies liegt daran, daß je nach Zuordnungsvariante im Mittel unterschiedlich stark beanspruchte Prozessoren durchlaufen werden. Diese Freiheitsgrade des Modells machen eine Interpretation der erhaltenen Ergebnisse schwierig und können gegebenenfalls andere Effekte überdecken. Deshalb wird ohne Beschränkung der Allgemeinheit die Zuordnung von Bearbeitungsschritten zu Prozessoren so eingeschränkt, daß unabhängig von der Zuordnungsvariante die Bearbeitungsschrittfolgen eines Lastszenarios immer die gleiche mittlere Durchlaufzeit durch einen Netzknoten erfahren. Dies wird dadurch erreicht, daß jeder Prozessor (für alle Lastszenarios) genau einer Klasse zugeordnet wird und Bearbeitungsschritte gleichwahrscheinlich auf alle Prozessoren einer einzigen Klasse verteilt werden. Darüberhinaus darf auf

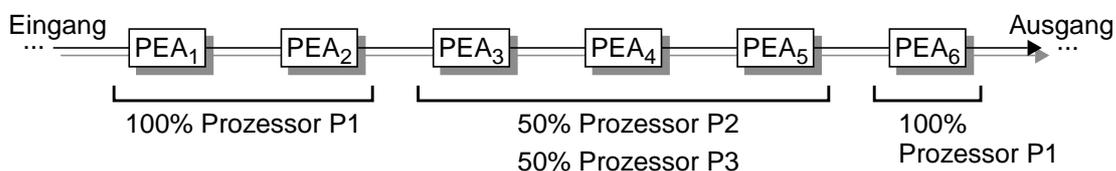


Abbildung 5-21: Beispiel einer eingeschränkten Zuweisung von Gruppen von Bearbeitungsschritten (*Physical Entity Actions, PEA*) an Prozessoren.

eine Zuweisung einer Gruppe von Bearbeitungsschritten eines Lastszenarios zu einer Klasse von Prozessoren keine Zuweisung der nachfolgenden Gruppe von Bearbeitungsschritten zu derselben Klasse von Prozessoren erfolgen.¹⁹ Abbildung 5-21 zeigt das Beispiel einer gültigen Zuweisung im Rahmen dieser Einschränkungen.

Ist eine Gruppe von Bearbeitungsschritten einem Prozessor zugewiesen, so konkurriert sie dort mit anderen Anforderungen um Bearbeitungszeit. Betrachtet man die einzelnen Bearbeitungsschritte, die im Rahmen des Lastmodells identifiziert wurden, so stellt man fest, daß es sich in der Regel um relativ einfache Vorgänge handelt, die einen Prozessor nur für kurze Zeit bean-

19. Geschieht dies doch, so kann es vorkommen, daß die Bearbeitung der ersten Gruppe im selben Prozessor erfolgt wie die der zweiten, was zu unterschiedlichen mittleren Durchlaufzeiten im Vergleich zu einer Auswahlvariante führt, bei der verschiedene Prozessoren beteiligt werden. Dies liegt daran, daß der Prozessor der ersten Gruppe zum Auswahlzeitpunkt gerade eine Bearbeitung abgeschlossen hat und deshalb andere mittlere Wartezeiten erlebt werden.

sprechen. Es macht daher Sinn, bei der Modellierung anzunehmen, daß während der Bearbeitung der Prozessor nicht mit anderen Anforderungen geteilt werden muß (kein „*processor sharing*“), sondern daß einem Bearbeitungsschritt der Prozessor in der Regel exklusiv bis zur vollständigen Abarbeitung zugeteilt ist. Eine Ausnahme von dieser Regel kann der Fall darstellen, in dem die Bearbeitung einer Anforderung unterbrochen wird, um Anforderungen höherer Priorität (vollständig) abzuarbeiten. Geht man von der beschriebenen Annahme aus, so bietet sich zur Modellierung eine Klasse von Bediensystemen an, die in [Pat90, PF89] detailliert untersucht und in [Bod94, BS97] erweitert wurden. Es handelt sich um Prioritätensysteme mit zufallsgesteuerten Verzweigungen, Aufspaltungen und Unterbrechungsdistanzen, bei denen Anforderungen eine endliche Folge von Bearbeitungsschritten gleicher oder unterschiedlicher Priorität durchlaufen können. Derartige Systeme wurden bereits in [Wil88, WK90, BKWZ92] verwendet, um die Bearbeitung von Protokollfunktionen des Signalisiersystems Nummer 7 zu modellieren.

Diese Arbeit beschränkt sich auf Bediensysteme, welche innerhalb einer Prioritätsstufe die Bedienreihenfolge FIFO einhalten und nach einer Unterbrechung die Bearbeitung einer Anforderung ohne Zeitverlust wieder aufnehmen. Darüberhinaus wird auf zufallsgesteuerte Verzweigungen verzichtet, da diese die Interpretation der Ergebnisse erschweren können. Alle diese Einschränkungen erfolgen ohne Beschränkung der Allgemeinheit. Das in dieser Arbeit verwendete Bediensystem läßt sich damit folgendermaßen charakterisieren:

- ❑ Jede Anforderung wird genau einer Klasse zugeordnet.
- ❑ Eine Anforderungsklasse ist durch eine Priorität, die Zufallsvariable Bearbeitungszeit und eine Beschreibung des Verhaltens nach Bedienung charakterisiert.
- ❑ Anforderungen können in Gruppen gleicher oder verschiedener Klassen auftreten.
- ❑ Es ist eine globale Unterbrechungsdistanz definiert.
- ❑ Trifft eine Anforderung auf eine leere Bedieneinheit oder auf eine Anforderung in Bedienung, deren Priorität mindestens um die Unterbrechungsdistanz niedriger ist, so wird sie sofort bedient. (Es sei denn, eine Anforderung höherer Priorität trifft gemeinsam mit ihr ein).
- ❑ Kann eine Anforderung nicht sofort bedient werden, so reiht sie sich entsprechend ihrer Priorität in die globale Warteschlange ein (FIFO mit Warteprioritäten).
- ❑ Nach Bearbeitung einer Anforderung kann diese je nach Klasse das System verlassen und/oder sich in eine oder mehrere Anforderungen neuer Klassen²⁰ verwandeln. Anschließend wird begonnen, diejenige Anforderung im System zu bearbeiten, welche die höchste Priorität hat.
- ❑ Die Unterbrechung und Wiederaufnahme der Bearbeitung von Anforderungen geschieht ohne zusätzliche Prozessorbelastung (ohne „*overhead*“) und ohne eine Veränderung der beanspruchten gesamten Bearbeitungsdauer („RESUME-Verfahren“).

20. Eine Anforderung kann sich nicht in eine Klasse verwandeln, die sie schon einmal durchlaufen hat, da sonst die Endlichkeit der Folge von Bearbeitungsschritten nicht mehr gewährleistet wäre.

- Jedem Prozessor wird zusätzlicher Verkehr („Störverkehr“) einer eigenen Klasse zugeordnet, der die Bearbeitung von Aufgaben modelliert, die nicht im Lastmodell enthalten sind.

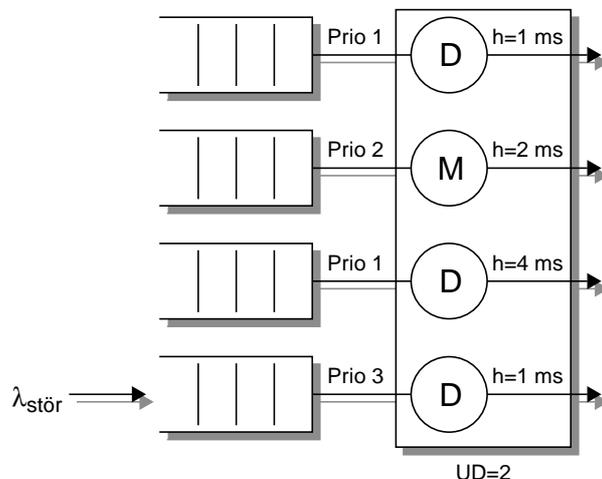


Abbildung 5-22: Beispiel eines Prozessormodells.

Üblicherweise wird solch ein Bediensystem so dargestellt, daß für jede Anforderungsklasse eine eigene (virtuelle) Warteschlange vorgesehen ist, die mit dem zugehörigen Prioritätswert markiert ist und mit einer Bedienphase innerhalb der Bedieneinheit verbunden ist. Tatsächlich warten die Anforderungen jedoch in einer globalen Warteschlange. Innerhalb dieser Warteschlange sind sie nach Prioritäten und bei gleicher Priorität nach dem FIFO-Prinzip geordnet. Jede Bedienphase wird mit den Charakteristika der zugehörigen Bedienzeit versehen. Die Bedieneinheit wird als Rechteck, das alle möglichen Bedienphasen umschließt, dargestellt. Die globale Unterbrechungsdistanz wird unter der Bedieneinheit notiert. Abbildung 5-22 zeigt ein entsprechendes Beispiel, bei dem überdies noch der prozessorspezifische „Störverkehr“ ($\lambda_{\text{stör}}$) – nicht jedoch Verkehr, der aus dem Lastmodell resultiert – eingezeichnet ist.

Das vorgestellte zweistufige Verfahren zur Prozessorenzuweisung stellt den Kern des Modells eines Netzknotens dar. Darüberhinaus können auch Verzögerungsglieder und Modelle eines Übertragungsabschnitts (siehe Abbildung 5-18) als Modellkomponenten zur Beschreibung der inneren Struktur eines Netzknotens verwendet werden.

5.5 Zusammenfassung

In diesem Kapitel wurden die Auslöser von Mobilitätsverwaltungsprozeduren, der Ablauf dieser Prozeduren und die sie ausführenden und unterstützenden Instanzen mit Hilfe eines Teilnehmermodells, eines Lastmodells und eines Netzmodells charakterisiert. Daraus läßt sich ermitteln, welche Mobilitätsverwaltungsprozeduren wie häufig angestoßen werden, welche Last an welchen Stellen des Netzes anfällt und in welcher Weise die anfallende Last verarbeitet wird. Im folgenden Kapitel wird auf Basis der so gewonnenen Informationen eine exemplarische Leistungsbewertung einiger Verfahren bei gegebener Netzstruktur vorgenommen.

Kapitel 6

Leistungsbewertung ausgewählter Beispiele

In diesem Kapitel wird exemplarisch gezeigt, wie basierend auf der im letzten Kapitel vorgestellten Modellbildung Aussagen über die Leistungsfähigkeit verschiedener Verfahren zur Aufenthaltsortsdatenverwaltung getroffen werden können. Die Fülle der Einsatzmöglichkeiten der vorgestellten Modellbildung und der begrenzende Rahmen der Arbeit legen eine Beschränkung bei der Auswahl der zu untersuchenden Systeme nahe. Es wird deshalb beispielhaft nur eine Netzstruktur vorgegeben, die ein relativ großes, hierarchisch gegliedertes Netz darstellen soll. Es wird nur eine kleine Gruppe möglicher Verfahren betrachtet, und auch aus dem Parameterraum des Teilnehmermodells wird nur ein kleiner, exemplarischer Teil untersucht. Bei der Methodik der Leistungsbewertung bleibt das Kapitel auf die analytische Bestimmung von Erwartungswerten beschränkt. Für eine Anwendung im Rahmen von simulativen Leistungsuntersuchungen ist die vorgestellte Modellbildung jedoch genau so gut geeignet.

Die Kriterien zur Leistungsbewertung lassen sich in zwei Gruppen einteilen. Die erste Gruppe bezieht sich auf die Ausführungskosten bzw. den Ressourcenbedarf verschiedener Mobilitätsverwaltungsprozeduren, was im folgenden auch als „Verkehrsangebot“ bezeichnet wird. Dazu gehört die Zahl und Größe von Signalisierungsmeldungen, die über Kommunikationsabschnitte transportiert werden müssen, sowie der Aufwand zur Verkehrsführung und vor allem zur Verarbeitung dieser Meldungen. Bei der Verarbeitung der Meldungen spielt der Rechenzeitbedarf in den verarbeitenden Prozessoren eine Rolle, ebenso wie der Zugriff auf den Datenbestand eines lokalen Datenbanksystems. Da die Größe des Datenbestands, die vom Verfahren zur Aufenthaltsortsdatenverwaltung abhängt, Einfluß auf den Ressourcenbedarf hat, wird auch sie zur Bewertung herangezogen. Die zweite Gruppe von Kriterien bezieht sich auf das Antwortzeitverhalten der Prozeduren. Während die erste Gruppe von Kriterien einem Netzbetreiber hilft, bei der Dimensionierung von Netzelementen eine Minimierung von Kosten zu erreichen, stellt die zweite Gruppe Kriterien zur Verfügung, die sich auf die vom Nutzer eines Kommunikationsdienstes empfundene Dienstgüte auswirken. In der Regel kann das Antwortzeitverhalten als eine einzuhaltende Randbedingung bei der Optimierung der Ausführungskosten betrachtet werden.

Die nachfolgenden Unterkapitel gliedern sich in drei Teile. Kapitel 6.1 beschreibt die Voraussetzungen für die nachfolgende Leistungsbewertung, insbesondere die untersuchte Netzstruktur, die Parameter des Teilnehmermodells sowie die exemplarisch ausgewählten Ver-

fahren zur Aufenthaltsortsdatenverwaltung. In Kapitel 6.2 werden die verschiedenen Verfahren und Konfigurationen in Bezug auf die Ausführungskosten verglichen. In Kapitel 6.3 geschieht dies auf der Basis von Antwortzeiten. Kapitel 6.4 faßt die Ergebnisse der beiden letzten Teile kurz zusammen.

6.1 Voraussetzungen

6.1.1 Untersuchte Netzstruktur und Parameter des Teilnehmermodells

Untersucht wird ein relativ großes Netz, das hierarchisch in drei Netzebenen gegliedert ist (siehe Abbildung 6-1). Jeder Knoten einer der beiden unteren Netzebenen ist mit genau einem Knoten in der jeweils nächsthöheren Netzebene („Vaterknoten“) verbunden. Die Knoten in den beiden oberen Netzebenen sind mit genau 16 Knoten in der jeweils nächstunteren Netzebene verbunden. Die oberste Netzebene besteht aus vier vollvermaschten Knoten. Die mittlere Netzebene besteht aus 64 Knoten, deren Verknüpfungen ein Netz mit quadratischen Maschen bilden, d.h. aus Knoten, die in einem quadratischen Raster angeordnet sind und die jeweils mit ihren – in der Regel vier – direkten Nachbarn verbunden sind. Die 1024 Knoten in der untersten Netzebene sind nicht direkt miteinander verknüpft.

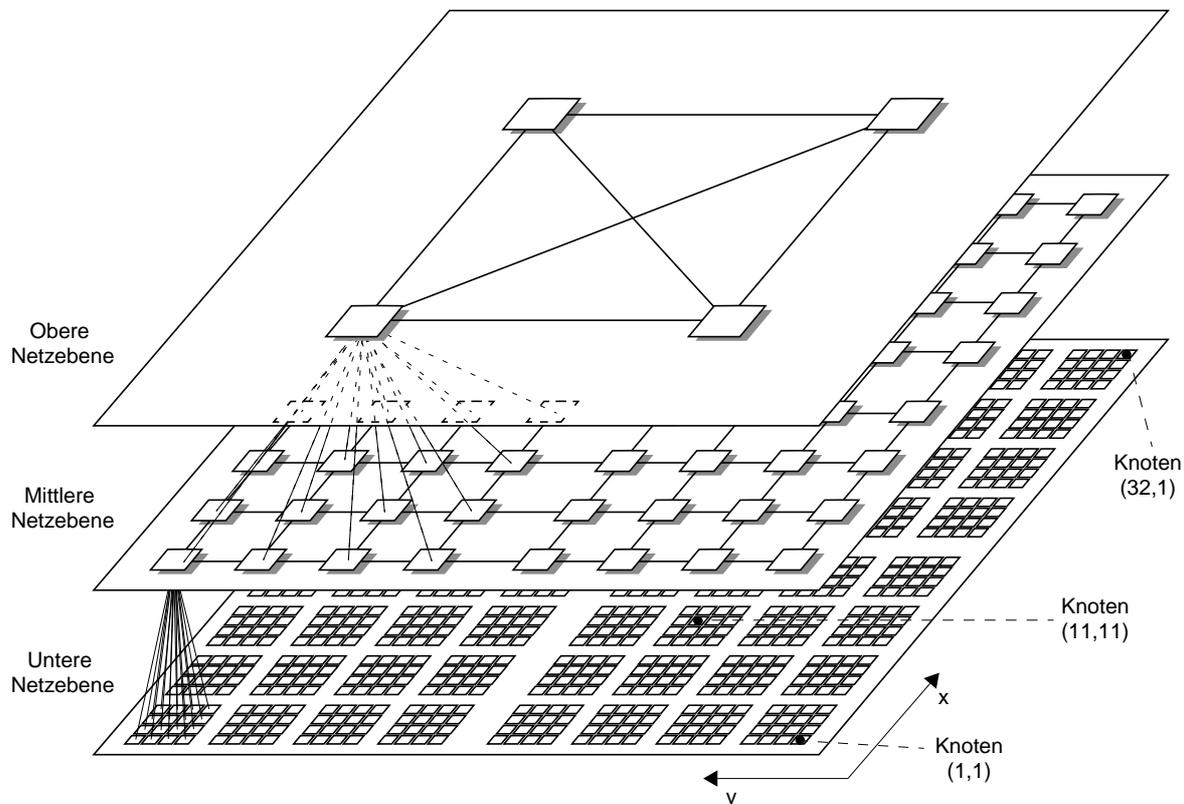


Abbildung 6-1: Struktur des untersuchten Netzes: 4 vollvermaschte Knoten in der oberen Netzebene, $4 \cdot 16 = 64$ Knoten in der mittleren Netzebene, die mit bis zu vier Nachbarknoten verbunden sind, und $64 \cdot 16 = 1024$ Knoten in der unteren Netzebene, die sternförmig an je genau einen Knoten der mittleren Ebene angebunden sind.

Teilnehmer bewegen sich ausschließlich zwischen Knoten der unteren Netzebene. Ebenso befinden sich Ursprung und Ziel von Rufen immer in solchen Knoten.¹ Bezeichnet man die Knoten der unteren Ebene mit den Koordinaten (x,y) ($x \in \{1, 2, \dots, 32\}$, $y \in \{1, 2, \dots, 32\}$), so kann man die Distanzen zwischen den Knoten proportional zu $\sqrt{(\Delta x)^2 + (\Delta y)^2}$ wählen. Die Distanzmatrix Δ ergibt sich dann nach (5-58) bis (5-60). Jeder Knoten der unteren Netzebene ist „Heimatknoten“ genau einer Teilnehmerklasse. Die 1024 Teilnehmerklassen k unterscheiden sich nur durch den Heimatknoten $h^{(k)}$, alle anderen Parameter des Teilnehmermodells werden jeweils identisch gewählt. Dazu gehört auch die Zahl der Teilnehmer je Klasse. An jedem Knoten der unteren Netzebene befindet sich eine Besucherdatei (VLR). Ortsaktualisierungen werden folglich durch den Wechsel des besuchten Netzknosens ausgelöst.

Das Bewegungsverhalten wird durch die Vorgabe der $\pi_u^{(k)} \sim f_\delta(d_{u, h^{(k)}})$ mit $f_\delta(d)$ entsprechend (5-61) charakterisiert, wobei im folgenden der Spreizfaktor δ zur Beschreibung des Bewegungsverhaltens mit δ_B (B wie Bewegung) bezeichnet wird. Daneben wird $\mathbf{P}^{(k)}$ durch die Gewichte $G_u^{(k)} = \pi_u^{(k)}$ nach (5-24) bestimmt, was eine sehr einfache Berechnung der $\phi_{u,v}^{(k)}$ nach (5-26) erlaubt. Aus den $\phi_{u,v}^{(k)}$ ergeben sich direkt die Raten der Aufenthaltsortsaktualisierungen.

Das Rufmodell wird ebenfalls in Bezug auf den Heimatknoten $h^{(k)}$ definiert. Auch hier wird der Spreizfaktor entsprechend verwendet und im folgenden als δ_R (R wie Rufe) bezeichnet. Die Ruftypen I und II werden im vorliegenden Beispiel im Verhältnis 1:1 angenommen. Die Rufraten ergeben sich aus den Gleichungen (5-51) und (5-62), (5-63) sowie (5-66), (5-67). Die Ruftypen III und IV werden nicht direkt betrachtet, sondern lediglich teilweise durch ein wachsendes δ_R für die Ruftypen I und II bei wachsendem δ_B berücksichtigt. In der Regel wird der Spreizfaktor der Rufe jedoch zu 0,05 gewählt. Für einen Teilnehmer der Klasse (11,11) beispielsweise bedeutet dies, daß Rufe vom Typ I zu rund 20% von seinem Heimatknoten, zu 52% von den restlichen Knoten mit dem selben Vaterknoten und zu 28% von allen anderen Knoten kommen. Dabei sind die Rufe von Knoten unterhalb der drei Knoten in der obersten Ebene, die nicht Vaterknoten des Vaterknotens sind, mit 0,3% fast vernachlässigbar.²

6.1.2 Die untersuchten Verfahren zur Aufenthaltsortsdatenverwaltung

Die Auswahl der verschiedenen Verfahren zur Aufenthaltsortsdatenverwaltung wurde so getroffen, daß sie als Fortentwicklung der bestehenden GSM-Mobilitätsverwaltung realisiert werden können. Bei allen Verfahren wird die Authentisierung von Teilnehmern und die Verwaltung von teilnehmerspezifischen Daten genau so wie bei GSM durchgeführt, d. h. es erfolgt eine funktionale Trennung zwischen HLR und VLR, Authentifizierungsparametersätze werden von VLR zu VLR weitergereicht und der Teilnehmerdatensatz wird bei jeder Aufenthaltsorts-

-
1. Bei diesem Beispiel wird ein geschlossenes System betrachtet. Bei einem realen Netz müßte auch Verkehr von und zu anderen Netzen berücksichtigt werden, der in der Regel über einige wenige Netzübergangsknoten geführt wird.
 2. Für $\delta_R = 0,1$ ergibt sich ungefähr 5% : 30% : 58% : 7%.

aktualisierung vom HLR in das VLR des neuen besuchten Knotens kopiert. Lediglich bezüglich der Ortsdatenverwaltung unterscheiden sich die untersuchten Verfahren. Es werden drei Verfahren verglichen:

- ❑ standardkonforme GSM-Mobilitätsverwaltung
- ❑ eine Erweiterung um *Proxy HLRs* (siehe Kapitel 4.2.4), bei der das GSM MAP-Protokoll beibehalten werden kann, und
- ❑ ein Ansatz, der zur Ortsdatenverwaltung ein hierarchisches, verteiltes Datenbanksystem verwendet, wie es in Kapitel 3.5.2 beschrieben ist.

Im Rahmen der GSM-Mobilitätsverwaltung gibt es Freiheitsgrade bezüglich der Platzierung von HLRs im Netz und der Zuordnung von Teilnehmern zu HLRs. Diese werden in Kapitel 6.2.1 näher untersucht.

Bei der Erweiterung um *Proxy HLRs* werden die HLRs in der oberen Netzebene und die *Proxy HLRs* in der mittleren Netzebene platziert. *Proxy HLRs* können entweder nur zur Signalisierungskostenreduktion bei Aufenthaltsortsaktualisierungen oder zusätzlich auch zur Kostenreduktion bei der Teilnehmersuche eingesetzt werden. Im letzteren Fall wird zuerst an einem Knoten der mittleren Netzebene überprüft, ob der gesuchte Teilnehmer dort registriert ist, und dann gegebenenfalls die Suche von dort direkt zum besuchten Knoten geführt. Im ersteren Fall geht die Ortsanfrage direkt an das HLR und wird von dort via das *Proxy HLR* an den aktuellen Aufenthaltsort geführt. In der Regel erfolgen die folgenden Untersuchungen getrennt für Aufenthaltsortsaktualisierungen (*Location Update*) und Teilnehmersuchen. Es wird davon ausgegangen, daß das Verfahren in beiden Fällen zur Signalisierungskostenreduktion eingesetzt wird.

Beim Verfahren mit einem hierarchischen, verteilten Datenbanksystem befinden sich die HLRs in der obersten Netzebene. Jeder Netzknoten beinhaltet daneben einen Datenbankknoten, d.h. die Netzhierarchie entspricht genau der Datenbankhierarchie. Die Datenbankhierarchie benötigt noch einen Wurzelknoten, der so in der Netzhierarchie nicht vorhanden ist. Dazu werden zwei Möglichkeiten untersucht: a) Das HLR eines Teilnehmers übernimmt die Rolle des Wurzelknotens und b) die vier Knoten der obersten Ebene realisieren einen verteilten Wurzelknoten. Ein verteilter Wurzelknoten bedeutet, daß jeder Knoten in der obersten Ebene dieselbe Information halten muß, die auch in einem einzelnen Wurzelknoten anzutreffen wäre. Die untersuchten Verfahren werden in den folgenden Diagrammen mit den Abkürzungen „PHLR“ für *Proxy HLR* und „HDB“ für die Ortsdatenverwaltung mit einem hierarchischen Datenbanksystem bezeichnet.

6.2 Vergleich auf der Basis des Verkehrsangebots

Im folgenden werden die beiden Mobilitätsverwaltungsprozeduren Teilnehmersuche und Aufenthaltsortsaktualisierung unabhängig voneinander betrachtet und nicht durch das Verhältnis ihres Auftretens miteinander in Beziehung gesetzt. Als sehr einfaches Vergleichskriterium soll zu Anfang lediglich die Zahl der „Hops“ je Prozedur dienen. Die Mobilitätsverwaltungsprotokolle basieren in der Regel auf entfernten Prozeduraufrufen. Die Entfernung zwischen aufrufendem Knoten und bearbeitendem Knoten soll die Basis der Bewertung sein. Dazu wird hier die Zahl der Kanten im Graphen der Netzstruktur als „Hop“ bezeichnet. Summiert man alle Entfernungen zwischen aufrufenden und bearbeitenden Instanzen je Mobilitätsverwaltungsprozedur auf, so erhält man für jede Prozedur die Zahl der „Hops“ und damit ein Maß für das Signalisieraufkommen im Netz. Das Vergleichskriterium bzw. die Kostenfunktion wird hier bewußt einfach gehalten, um die Parameter des Systems und die Ergebnisse überschaubar zu halten. Ebenso gut könnte jedoch auch eine komplexere Kostenfunktion verwendet werden, die beispielsweise die Kosten für die Übertragung von Signalisierungsmeldungen mit der Entfernung zwischen Knoten korreliert oder die in Abhängigkeit von der Netzebene die Kosten je Hop unterschiedlich annimmt, so daß beispielsweise Signalisierung in den oberen Netzebenen auf Grund von gebündelterem Verkehr mit weniger Kosten belegt werden könnte.

6.2.1 Platzierung von Netzknoten bei GSM-basierter Mobilitätsverwaltung

Das HLR eines Teilnehmers wird in GSM durch *Global Titles* im Rahmen des Signalisierungssystems Nummer 7 adressiert. Verwendung finden dabei die Internationale Mobilteilnehmerkennung (IMSI) sowie die öffentlichen Teilnehmerrufnummer (MSISDN). In der Regel wird die IMSI bei Aufenthaltsortsaktualisierungen verwendet, die MSISDN dagegen bei der Teilnehmersuche. Um die Tabellen bei der *Global Title Translation* klein zu halten, werden die beiden Kennzeichner aus Nummernblöcken entnommen, die administrativ jeweils dem selben HLR zugeordnet sind. Eine Zuordnung von Teilnehmer zu HLR kann dabei von verschiedenen Kriterien geprägt sein. Beispielsweise können Nummernblöcke und damit auch die zuständigen HLRs zwischen verschiedenen Diensteanbietern verteilt werden. Umgekehrt kann durch eine „Wunschrufnummer“ die Zuordnung zu einem bestimmten HLR festgelegt sein. Will man dagegen den Aufwand zur Ortsdatenverwaltung klein halten, so ordnet man einem wenig mobilen Teilnehmer ein HLR zu, das sich signalisieretechnisch relativ nahe an seinem Heimatknoten befindet. Diese Vorgehensweise wird in diesem Kapitel als „optimierte“ GSM-Mobilitätsverwaltung bezeichnet. Bei einer derart optimierten Zuordnung von Teilnehmern zu HLRs bietet es sich an, die Position von HLRs im Netz an das Mobilitätsverhalten der Teilnehmer anzupassen.

Die Abbildungen 6-2 und 6-3 zeigen die mittlere Zahl an „Hops“ für *Location Updates* bzw. für Teilnehmersuchen über dem Spreizfaktor δ_B der Bewegung für verschiedene Platzierungen von HLRs und unter Berücksichtigung der optimierten Zuordnung von Teilnehmern zu HLRs. Gewöhnliches GSM, bei dem die Teilnehmer zufällig auf die vorhandenen HLRs verteilt sind,

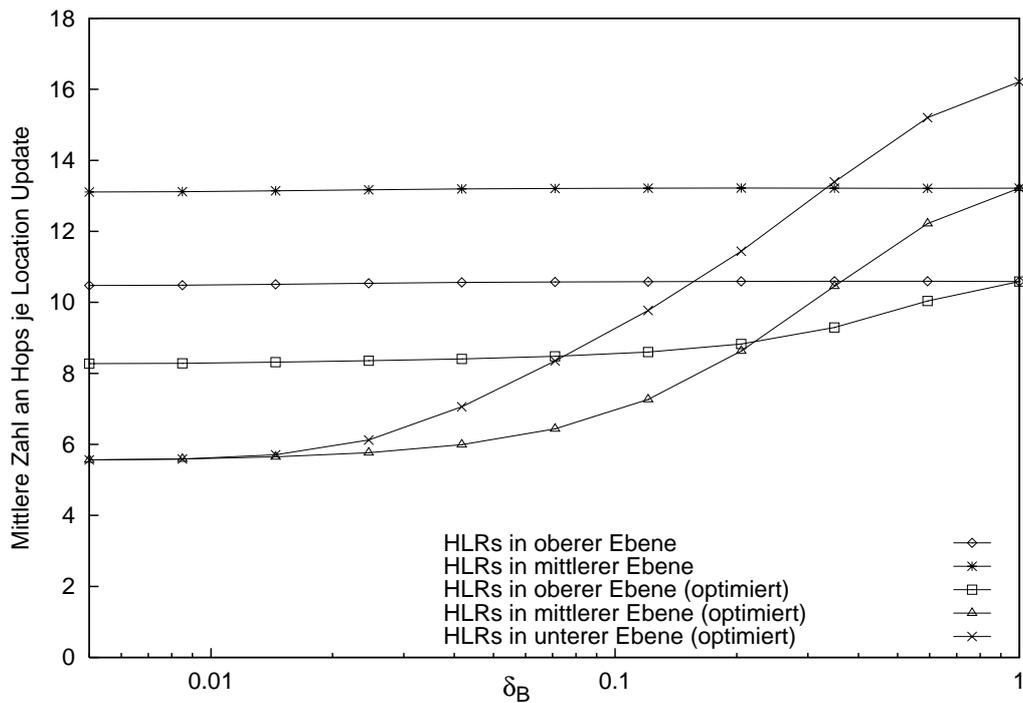


Abbildung 6-2: Vergleich der mittleren Zahl an „Hops“ je *Location Update* für verschiedene Möglichkeiten, GSM-Heimatdateien (HLRs) im Netz zu platzieren und Teilnehmer den Heimatdateien zuzuordnen.

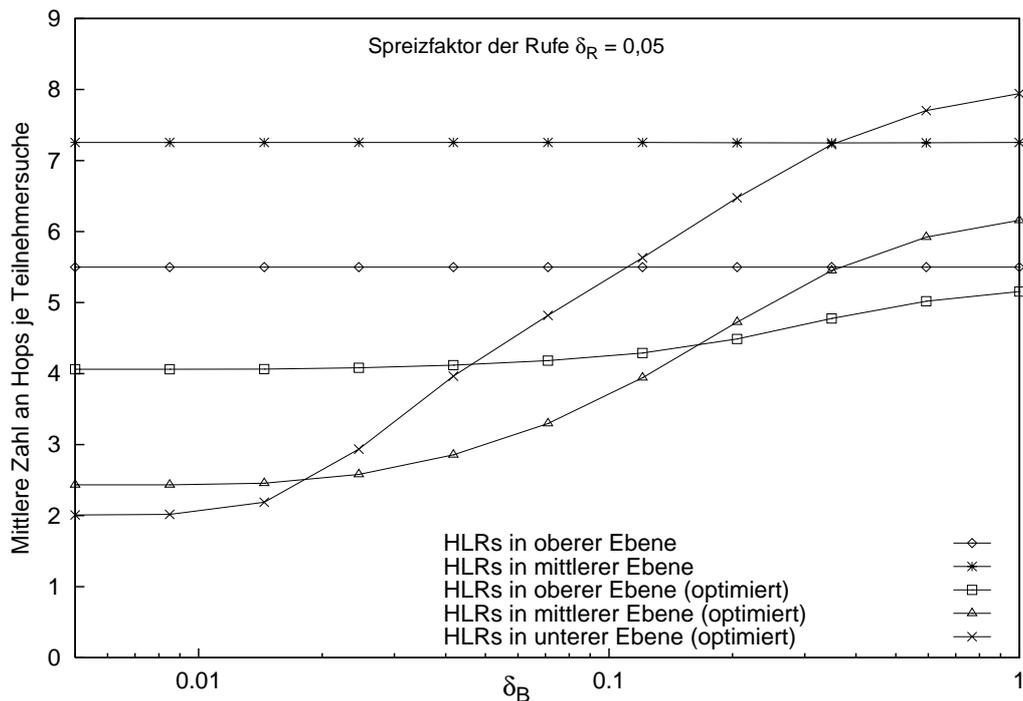


Abbildung 6-3: Vergleich der mittleren Zahl an „Hops“ je Teilnehmersuche für verschiedene Möglichkeiten, GSM-Heimatdateien (HLRs) im Netz zu platzieren und Teilnehmer den Heimatdateien zuzuordnen.

arbeitet am effizientesten, wenn alle (vier) HLRs in der oberen Netzebene platziert werden (\diamond). Platziert man dagegen alle (64) HLRs in der mittleren Netzebene ($*$), erhöhen sich die mittleren Kosten für beide Prozeduren signifikant. Dabei profitiert zwar unter Umständen ein Teil

der Teilnehmer von näheren HLRs, ein sehr viel größerer Teil wird jedoch von den nun weiter entfernten HLRs beeinträchtigt. Erst bei der optimierten Zuordnung von Teilnehmern zu HLRs kann sich das Plazieren von HLRs unterhalb der oberen Ebene lohnen. Bereits bei der Platzierung von HLRs in der oberen Ebene (\square) ergibt sich über den gesamten Mobilitätsbereich³ durch die Optimierung eine Verbesserung der Effizienz. Eine Platzierung in der mittleren Ebene (\triangle) bringt dann für kleine bis mittlere geographische Mobilität weitere deutliche Vorteile, wird jedoch bei höherer Mobilität sogar schlechter als nicht-optimiertes GSM mit vier HLRs in der oberen Ebene. Die Platzierung der HLRs in der unteren Ebene (\times) dagegen ist nur für kleine δ_B bei Teilnehmersuchen besser als die Platzierung in der mittleren Ebene. Ansonsten ergeben sich für diese Platzierung mit wachsendem δ_B immer schlechtere Werte. Verändert sich das Verhalten der Teilnehmer so, daß eine neue Zuordnung zu optimalen HLRs gewählt werden müßte, oder entspricht die Zuordnung zu HLR von Anfang an nicht dem Optimum, so können sich die Kosten für die Mobilitätsprozeduren bei der „optimierten“ Variante deutlich erhöhen. Für eine Platzierung der HLRs in der mittleren Netzebene deutet die Kurve für nicht-optimiertes GSM (\ast) an, wie hoch in solch einem Fall die Zusatzkosten im Mittel werden können.

6.2.2 Vergleich mit alternativen Verfahren

Die Verwendung eines hierarchischen, verteilten Datenbanksystems (HDB), wie es in Kapitel 3.5.2 beschrieben ist, hängt stark von der Definition eines „Heimatknotens“ ab. Eine Kette von Zeigern führt vom Heimatknoten zum aktuell besuchten Knoten. Je näher ein Teilnehmer sich an seinem Heimatknoten befindet, um so weniger Datenbankeinträge sind vonnöten. Stimmt die Definition des Heimatknotens nicht mit dem Teilnehmerverhalten überein, so resultiert dies zwar nicht in einer höheren Zahl an „Hops“, jedoch in einer höheren Zahl an Einträgen in den Datenbankknoten, insbesondere im Wurzelknoten. Dies macht das Verfahren unflexibel bei der Vergabe von „Wunschrufnummern“ und anfällig für Änderung des Teilnehmerverhaltens. Wie oben beschrieben, wird das Verfahren nur für die Ortsdatenverwaltung eingesetzt. Ansonsten werden HLRs und VLRs wie bei GSM verwendet. Die HLRs befinden sich in der obersten Ebene und die Zuordnung von Teilnehmern zu HLRs ist optimiert. Letzteres ist eine sinnvolle Annahme, da sich die Definition des Heimatknotens in der Vergabe von IMSI und MSISDN widerspiegelt, wodurch eine einfache Zuordnung zum optimalen HLR quasi automatisch gegeben ist. Beim Vergleich mit GSM, sollte deshalb beachtet werden, daß das Verfahren nur mit den „optimierten“ GSM-Varianten vergleichbar ist.

Die Abbildungen 6-4 und 6-5 vergleichen das alternative Verfahren (HDB) mit optimiertem GSM bezüglich der mittleren Zahl an „Hops“ über dem Spreizfaktor δ_B bei *Location Updates* bzw. bei Teilnehmersuchen. Es zeigt sich, daß die HDB-basierte Mobilitätsverwaltung (\times bzw. $+$) deutlich bessere Werte erzielt als „optimiertes“ GSM mit HLRs in der oberen Netzebene (\square). Außerdem ergibt sich, daß sie für geringe und mittlere Mobilität mit „optimiertem“ GSM,

3. Der Mobilitätsbereich wird bis $\delta_B = 1.0$ untersucht. Darüber wäre ein Teilnehmer seltener am Heimatknoten als an anderen Knoten, was die Definition „Heimatknoten“ sinnlos werden läßt.

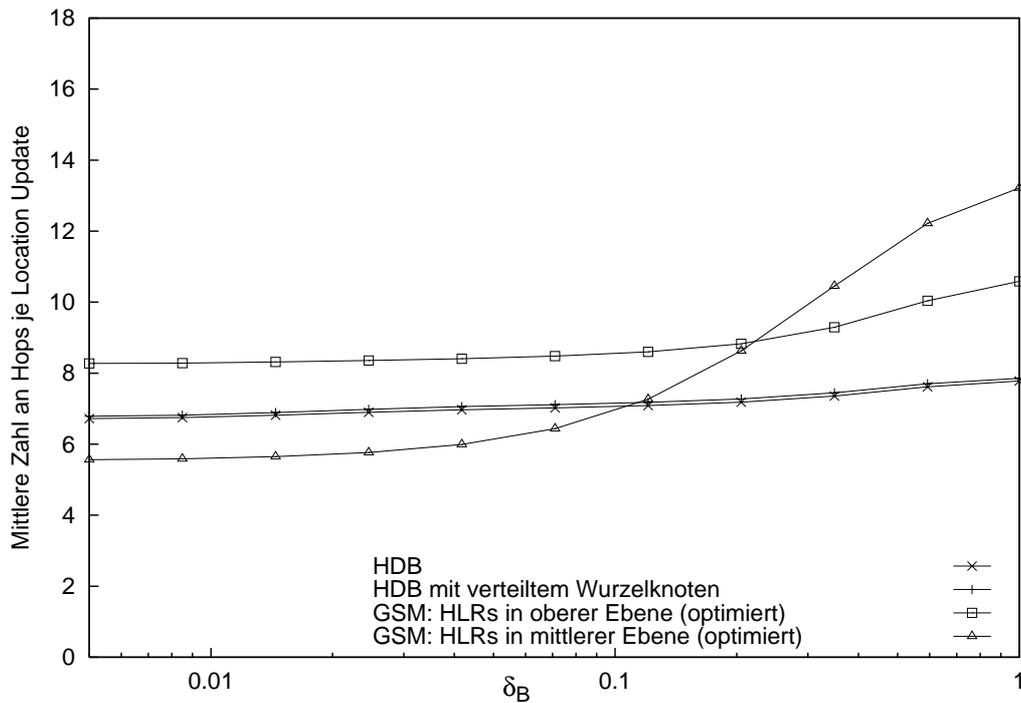


Abbildung 6-4: Vergleich zwischen HDB-basierter Mobilitätsverwaltung und verschiedenen Konfigurationen der GSM-Mobilitätsverwaltung bezüglich der mittleren Zahl an „Hops“ je *Location Update*.

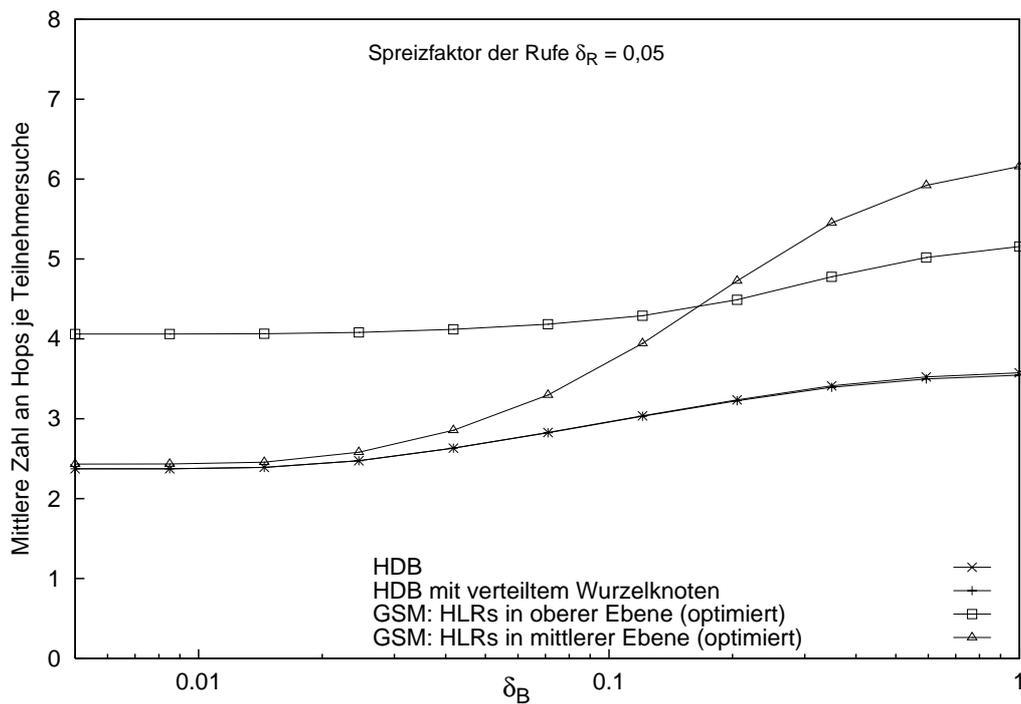


Abbildung 6-5: Vergleich zwischen HDB-basierter Mobilitätsverwaltung und verschiedenen Konfigurationen der GSM-Mobilitätsverwaltung bezüglich der mittleren Zahl an „Hops“ je Teilnehmersuche.

dessen HLRs sich in der mittleren Netzebene befinden (\triangle), vergleichbar ist und für höhere Mobilität sogar deutlich weniger „Hops“ benötigt. Der Vergleich zwischen dem HDB-Verfahren, bei dem die HLRs Wurzelknoten spielen (\times), und dem selben Verfahren, bei dem ein ver-

teilter Wurzelknoten vorliegt (+), ergibt, daß bei den gewählten Parametern für δ_B und δ_R die (geringen) zusätzlichen Kosten bei *Location Updates* beim Verfahren mit verteiltem Wurzelknoten nicht durch entsprechende Einsparungen bei der Teilnehmersuche gerechtfertigt werden können. (Die beiden Kurven in Abbildung 6-5 sind praktisch deckungsgleich).

Der in Kapitel 4.2.4 gemachte Vorschlag zur Verwendung eines *Proxy HLR* ist eine Erweiterung des standardisierten GSM-Mobilitätsverwaltungsverfahrens. Daher bestehen hier die gleichen Optionen bezüglich der HLR-Plazierung und der Zuordnung von Teilnehmern zu HLRs. An dieser Stelle soll jedoch nur die Plazierung der HLRs in der oberen Netzebene betrachtet werden.⁴ Unterschieden wird lediglich zwischen einer zufälligen Zuordnung von Teilnehmern zu HLRs („PHLR“) und einer optimierten Zuordnung („PHLR (optimiert)“).

In den Abbildungen 6-6 und 6-7 werden diese beiden Möglichkeiten mit gewöhnlichem GSM und mit der HDB-basierten Mobilitätsverwaltung verglichen. Bemerkenswert ist, daß das PHLR-Verfahren mit zufälliger Zuordnung von Teilnehmern zu HLRs (\square) – und damit mit entsprechender Freiheit bei der Vergabe von Rufnummer und Teilnehmerkennung – deutlich besser abschneidet als die entsprechende GSM-Konfiguration (–). Darüberhinaus ist das Verfahren nicht sehr viel schlechter als das HDB-basierte Verfahren (\times) und als das „optimierte“ PHLR-Verfahren (+). Beim direkten Vergleich der beiden letztgenannten Verfahren ist das „optimierte“ PHLR-Verfahren (+) für *Location Updates* nahezu gleich gut wie das HDB-basierte Verfahren (\times), während es bei der Teilnehmersuche geringfügig schlechter ist.

Zusammenfassend kann man sagen, daß bei beliebiger Zuordnung von Teilnehmern zu HLRs das PHLR-Verfahren die erste Wahl ist, und daß es auch bei optimierter Zuordnung von Teilnehmern zu HLRs gegenüber dem HDB-basierten Verfahren konkurrenzfähig ist. Für eine Strategie, bei der ein Teil der Teilnehmer optimalen HLRs zugeordnet wird, während ein anderer Teil mit „Wunschrufnummern“ versehen wird bzw. bei Veränderung des Bewegungsverhaltens (z.B. durch Wechsel des Heimatknotens) nicht mehr dem optimalen HLR zugeordnet ist, bietet sich das PHLR-Verfahren als eine sehr gute Hybrid-Lösung an

Die bisher präsentierten Ergebnisse setzen sich aus der Mittelwertbildung über alle Teilnehmerklassen zusammen. Für einzelne Klassen ergeben sich dagegen leicht abweichende Resultate. Besonders ausgeprägt ist dies für Klassen, deren Heimatknoten sich in der Mitte der unteren Netzebene befindet. Der Ablauf von Mobilitätsverwaltungsprozeduren resultiert dort häufig in relativ hohen Kosten, da die benachbarten Knoten im Baum der Netzstruktur weit voneinander entfernt sind. Dieser Effekt wirkt sich besonders bei kleinem δ_B aus, da dort sehr häufig Prozeduren zwischen dem Heimatknoten und den acht direkt benachbarten Netzknoten stattfinden. So kommt es, daß teilweise die Kosten für kleine δ_B größer sind als für größere δ_B . Die Abbildungen 6-8 und 6-9 illustrieren den Effekt für die Teilnehmerklasse (16,16).

4. Betrachtet man die guten Ergebnisse für HLRs in der mittleren Netzebene bei „optimiertem“ GSM für niedriger und mittlerer Mobilität, so bietet sich für diesen Mobilitätsbereich natürlich auch an, die HLRs beim „optimierten“ PHLR-Verfahren in der mittleren Netzebene zu plazieren.

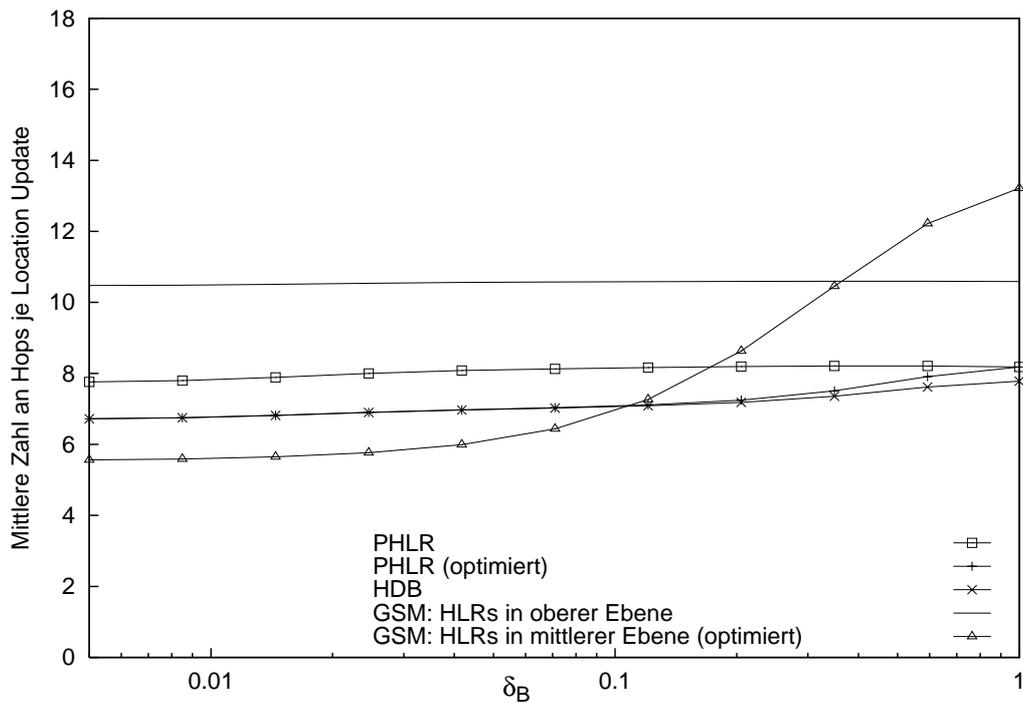


Abbildung 6-6: Vergleich zwischen einer *Proxy-HLR*-basierter Mobilitätsverwaltung und HDB- bzw. GSM-basierter Mobilitätsverwaltung bezüglich der mittleren Zahl an „Hops“ je *Location Update*.

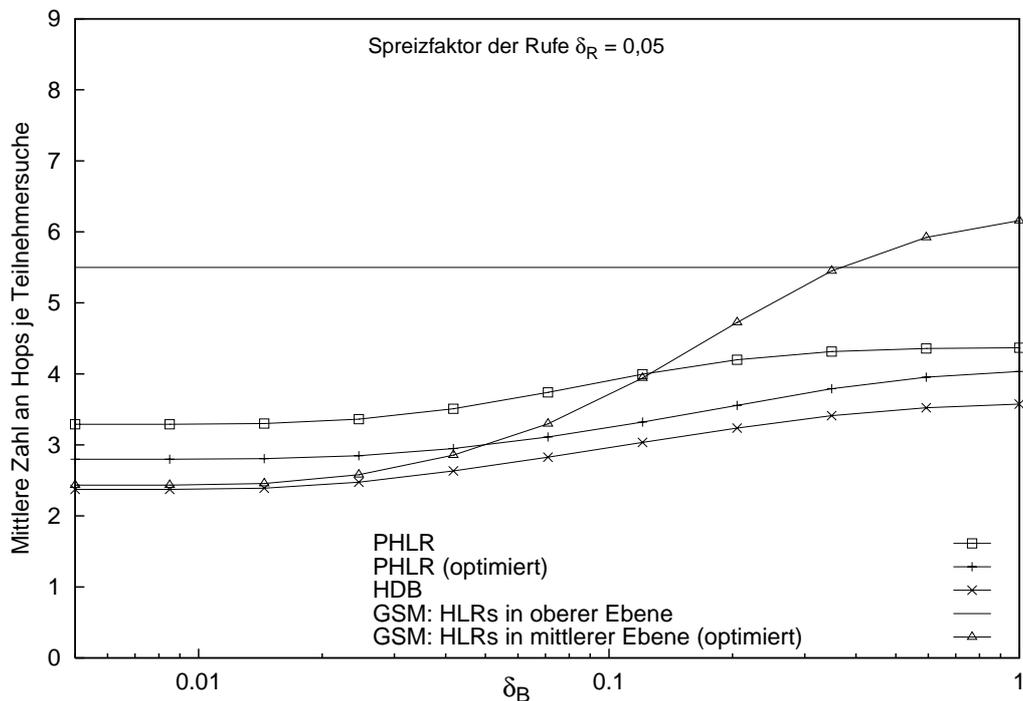


Abbildung 6-7: Vergleich zwischen einer *Proxy-HLR*-basierter Mobilitätsverwaltung und HDB- bzw. GSM-basierter Mobilitätsverwaltung bezüglich der mittleren Zahl an „Hops“ je Teilnehmersuche.

Als Vergleich dazu sind im Anhang A.2 die entsprechenden Ergebnisse für die Teilnehmerklasse (1,1) in den Abbildungen A-2 und A-3 dargestellt. Durch ihre Randlage entstehen für diese Klasse bei kleinem δ_B nur geringe Kosten für *Location Update* und Teilnehmersuche.

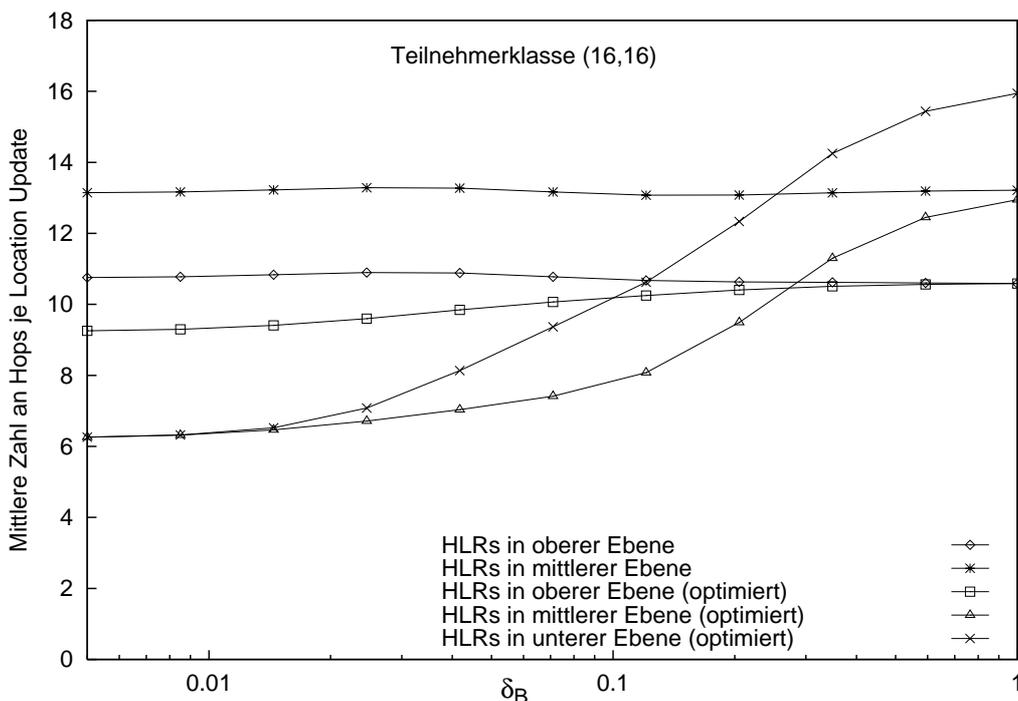


Abbildung 6-8: Gleiche Darstellung wie in Abbildung 6-2 für die Teilnehmerklasse mit Heimatknoten (16,16).

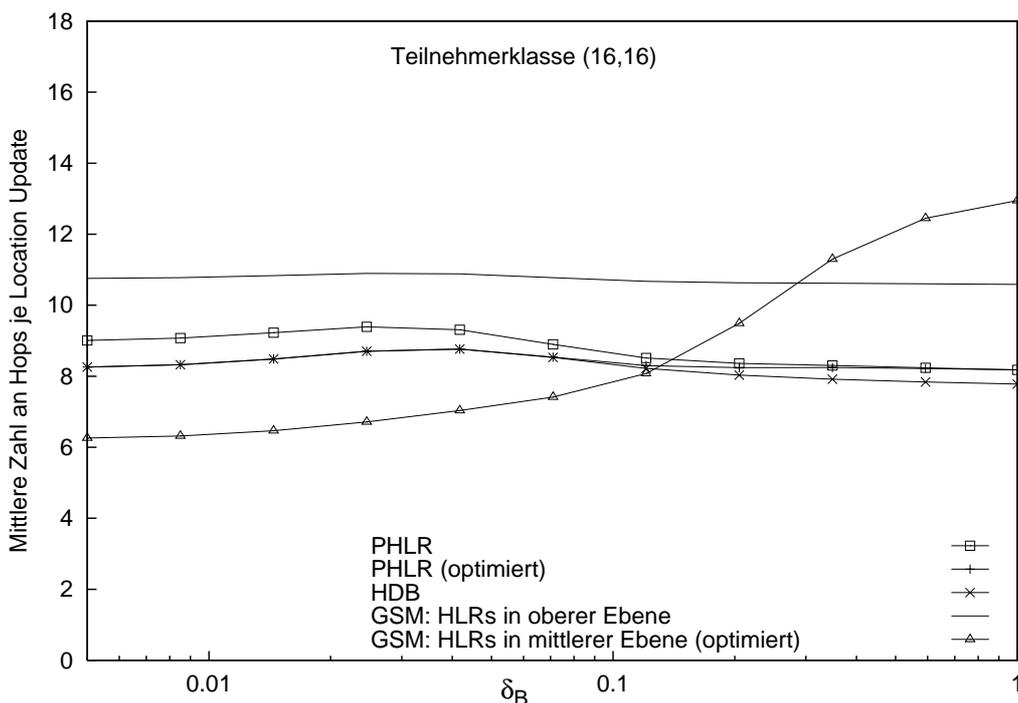


Abbildung 6-9: Gleiche Darstellung wie in Abbildung 6-6 für die Teilnehmerklasse mit Heimatknoten (16,16).

Die Überlagerung des Verhaltens aller 1024 Klassen führt zu den oben präsentierten Ergebnissen. Zur Verringerung des numerischen Aufwands zur Berechnung solcher Ergebnisse wird weiter unten teilweise die Teilnehmerklasse (11,11) als „repräsentative Klasse“ herangezogen. Abbildungen A-4 und A-5 dokumentieren die zugehörigen Ergebnisse.

6.2.3 Weitere Einflüsse auf die Kosten einer Teilnehmersuche

Ein wesentlicher Unterschied zwischen der HDB-basierten Mobilitätsverwaltung und den anderen untersuchten Verfahren besteht darin, daß im HDB-basierten System bei der Teilnehmersuche sofort erkannt wird, ob der gerufene Teilnehmer sich am Ursprungsort des Rufes befindet. In Abbildung 6-7 beispielsweise läßt sich der Unterschied zwischen dem optimierten PHLR-Verfahren und dem HDB-basierten Verfahren im wesentlichen durch diesen Effekt erklären. Am Ende von Kapitel 4.2.1 wurde bereits erwähnt, daß Signalisierung eingespart werden kann, indem vor einer entfernten Zielsuche zuerst das lokale VLR befragt wird, ob der Teilnehmer zur Zeit dort eingebucht ist. Dies soll hier als „VLR-Lookup“ bezeichnet werden. Abbildungen 6-10 und 6-11 zeigen, daß alle Verfahren (außer dem HDB-basierten) bei $\delta_R = 0,05$ von solch einem „VLR-Lookup“ signifikant profitieren. In den Diagrammen ist jedes Verfahren mit und ohne „VLR-Lookup“ dargestellt. In beiden Fällen wird jeweils das selbe Symbol verwendet, die Kurve für „VLR-Lookup“ verläuft aber unterhalb der jeweils anderen und ist gestrichelt dargestellt. In Abbildung 6-11 sieht man gut, daß die Kurve für das optimierte PHLR-Verfahren mit „VLR-Lookup“ (+, gestrichelt) über weite Bereiche mit der Kurve für das HDB-basierte Verfahren (–) fast deckungsgleich ist.

Dieses Ergebnis hängt bei dem hier verwendeten Teilnehmermodell natürlich stark von δ_R ab. Bereits bei $\delta_R = 0,1$ verschwinden die Vorteile des „VLR-Lookup“ fast vollständig. Auch die Unterschiede zwischen dem optimierten PHLR-Verfahren und dem HDB-basierten Verfahren sind für kleine und mittlere δ_B nahezu vernachlässigbar (siehe Abbildung A-6 im Anhang A.2).

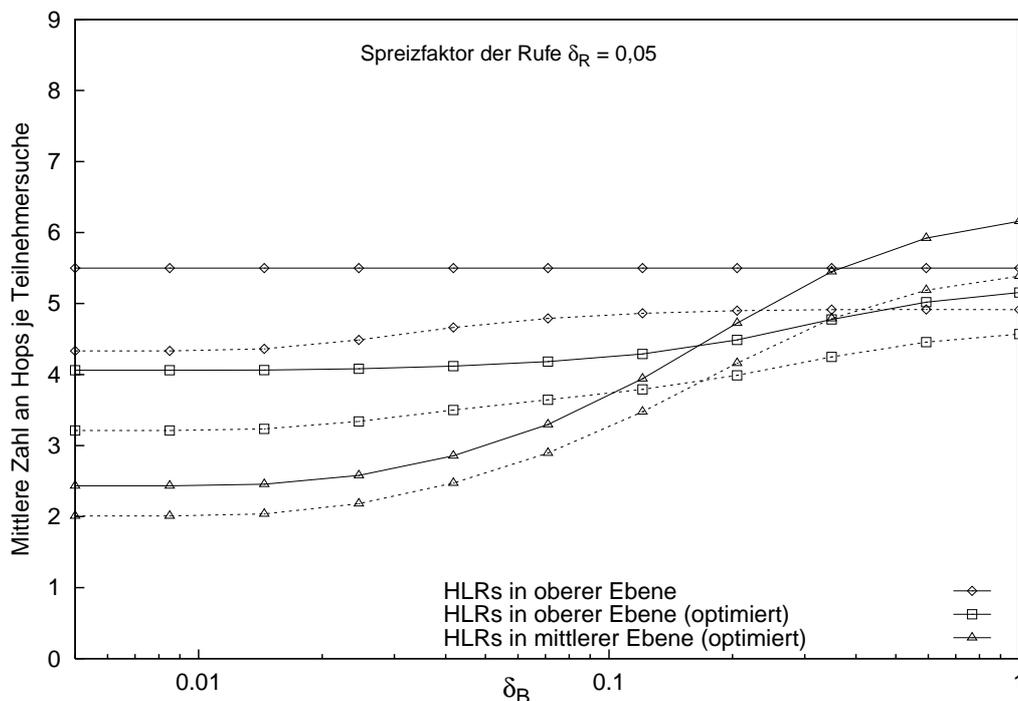


Abbildung 6-10: Vergleich zwischen GSM mit „VLR-Lookup“ (gestrichelt) und GSM ohne „VLR-Lookup“ (durchgezogen) bezüglich der mittleren Zahl an „Hops“ je Teilnehmersuche mit $\delta_R = 0,05$.

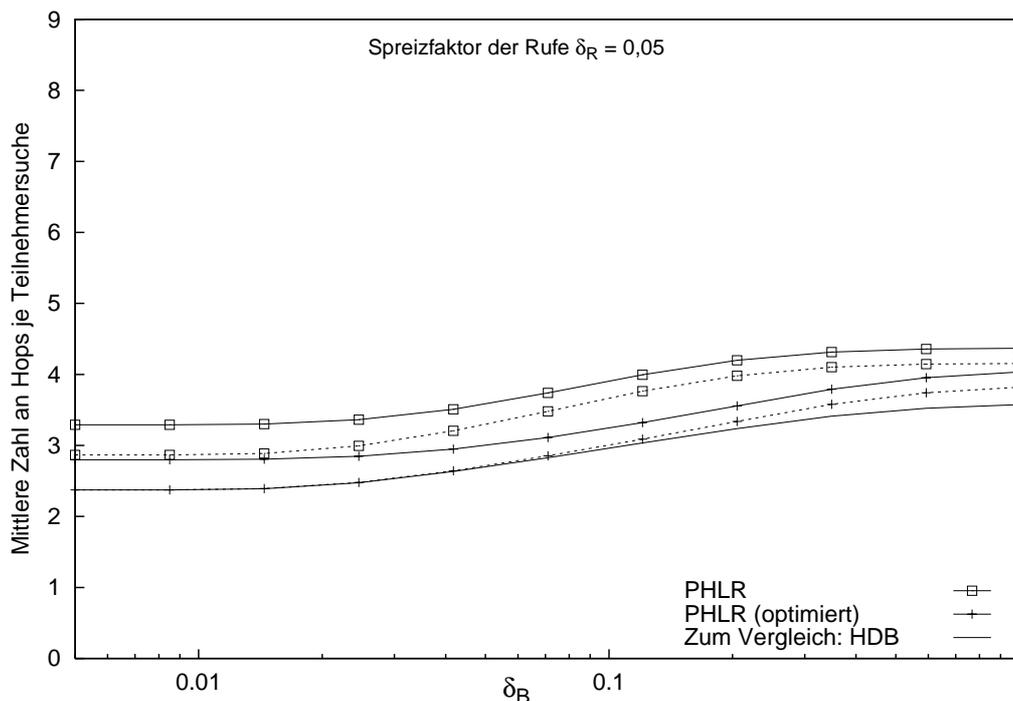


Abbildung 6-11: Vergleich zwischen dem *Proxy-HLR*-Verfahren mit „VLR-Lookup“ (gestrichelt) und demselben Verfahren ohne „VLR-Lookup“ (durchgezogen) bezüglich der mittleren Zahl an „Hops“ je Teilnehmersuche mit $\delta_R = 0,05$.

Die Abhängigkeit von der Lokalität der Rufe gilt natürlich für alle Verfahren außer dem nicht-optimierten GSM-Verfahren. Abbildungen A-7 und A-8 im Anhang A.2 zeigen, daß sich für alle Verfahren bezüglich der Teilnehmersuche bei $\delta_R = 0,1$ schlechtere Werte ergeben als bei $\delta_R = 0,05$. Dies gilt besonders für das HDB-Verfahren und das PHLR-Verfahren. Entsprechend erhöhen sich die Kosten einer Teilnehmersuche bei $\delta_R = 1,0$ weiter (siehe Abbildungen A-9 und A-10). Um diese Effekte noch einmal zu illustrieren und die Auswirkung von Rufen der Rufotypen III und IV abschätzen zu können, wird in den Abbildungen 6-12 und 6-13 der Spreizfaktor δ_R der Rufe zwischen 0,05 und 1,0 gemäß $\delta_R = 0,05 + \delta_B / 0,95$ variiert. In Abbildung 6-12 zeigt sich, daß optimiertes GSM mit den HLRs in der oberen Netzebene (□) nicht übermäßig beeinträchtigt wird und sich für große δ_B bzw. δ_R den Kosten für nicht-optimiertes GSM mit HLRs in der oberen Netzebene (—) annähert. Optimiertes GSM mit den HLRs in der mittleren Netzebene (△) dagegen ist sehr sensitiv auf wachsendes δ_R . Die Kosten für große δ_B bzw. δ_R übersteigen die Kosten von gewöhnlichem, nicht-optimiertem GSM deutlich. Abbildung 6-13 zeigt, daß sowohl das HDB-basierte Verfahren (× bzw. +) als auch das PHLR-Verfahren (□) höhere Kosten bei der Teilnehmersuche für höhere δ_B bzw. δ_R erfahren. Bei beiden Verfahren übersteigt jedoch die mittlere Zahl an „Hops“ nicht die von nicht-optimiertem GSM mit HLRs in der oberen Netzebene (—). Für $\delta_B = \delta_R = 1,0$ kommen sie dieser aber recht nahe. Anzumerken ist schließlich noch, daß für hohe δ_B bzw. δ_R das HDB-basierte Verfahren mit verteiltem Wurzelknoten (+) hier nun einen signifikanten Vorteil gegenüber dem normalen HDB-Verfahren hat (×).

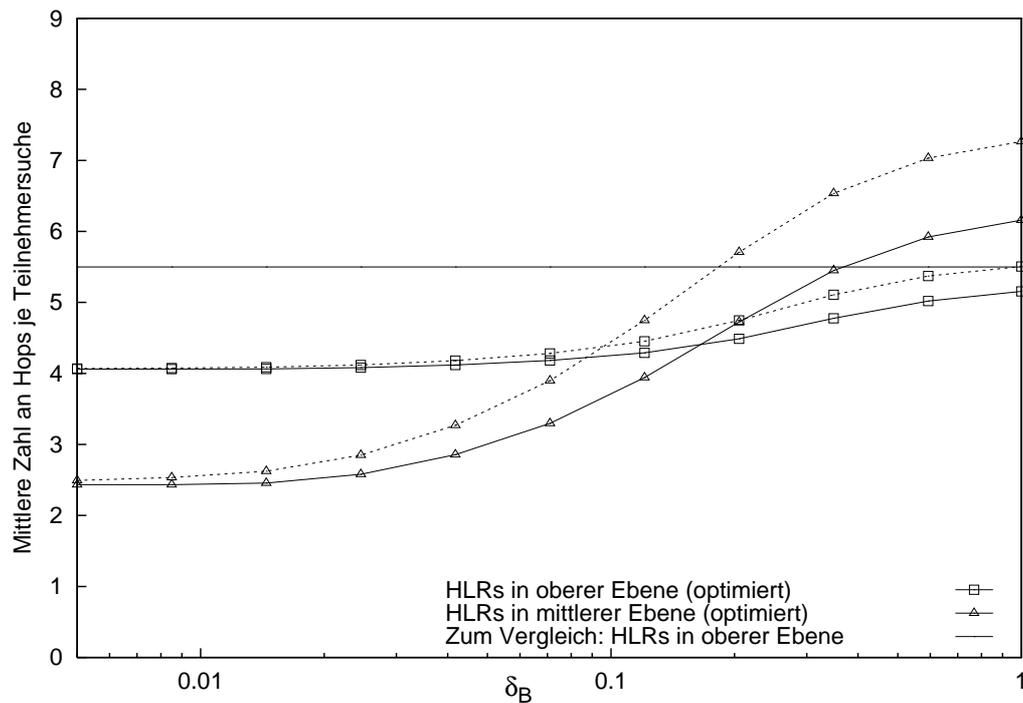


Abbildung 6-12: Mittlere Zahl an „Hops“ je Teilnehmersuche bei verschiedenen GSM-Konfigurationen für $\delta_R = 0,05$ (durchgezogen) und $\delta_R = 0,05 + \delta_B/0,95$ (gestrichelt).

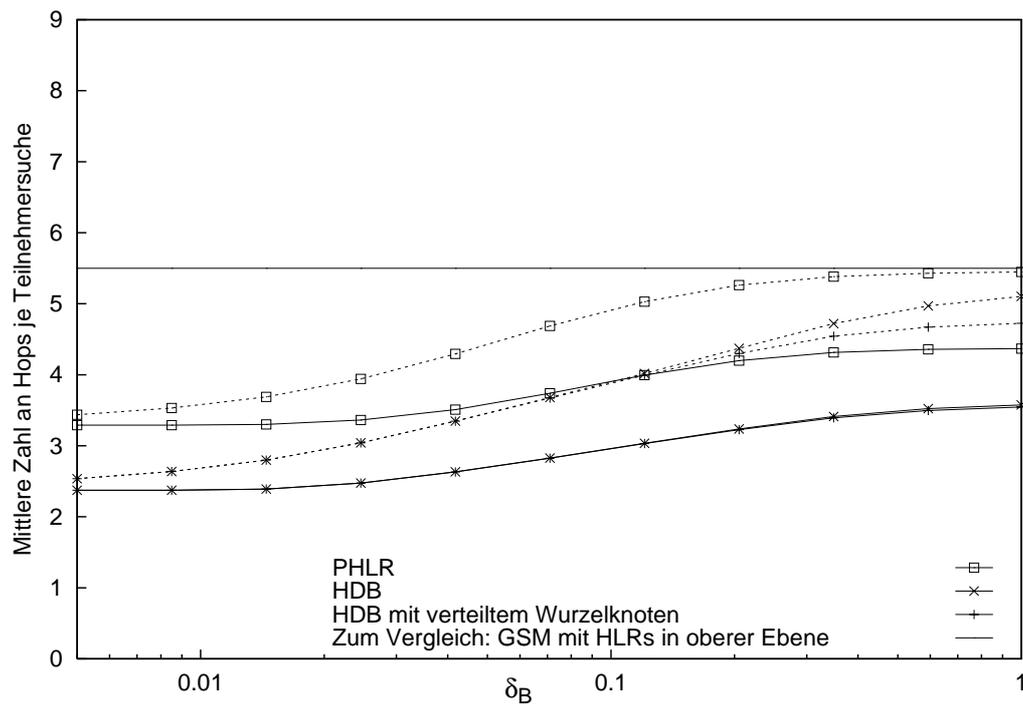


Abbildung 6-13: Mittlere Zahl an „Hops“ je Teilnehmersuche bei verschiedenen Verfahren zur Aufenthaltsortsdatenverwaltung für $\delta_R = 0,05$ (durchgezogen) und $\delta_R = 0,05 + \delta_B/0,95$ (gestrichelt).

6.2.4 Zahl der Datenbankeinträge und Datenbankzugriffe

In den vorhergehenden Unterkapiteln werden die verschiedenen Verfahren im wesentlichen bezüglich des zu erwartenden Signalisierverkehrs verglichen. Hier werden nun die Anforderungen an das Datenbanksystem als Vergleichskriterium herangezogen. Es zeigt sich, daß Einsparungen bei der Signalisierung teilweise durch erhöhten Aufwand im Datenbanksystem erkauft werden. Dieser erhöhte Aufwand stellt sich in Form einer größeren Zahl an Datenbankeinträgen und in einer größeren Zahl von Datenbankzugriffen dar.

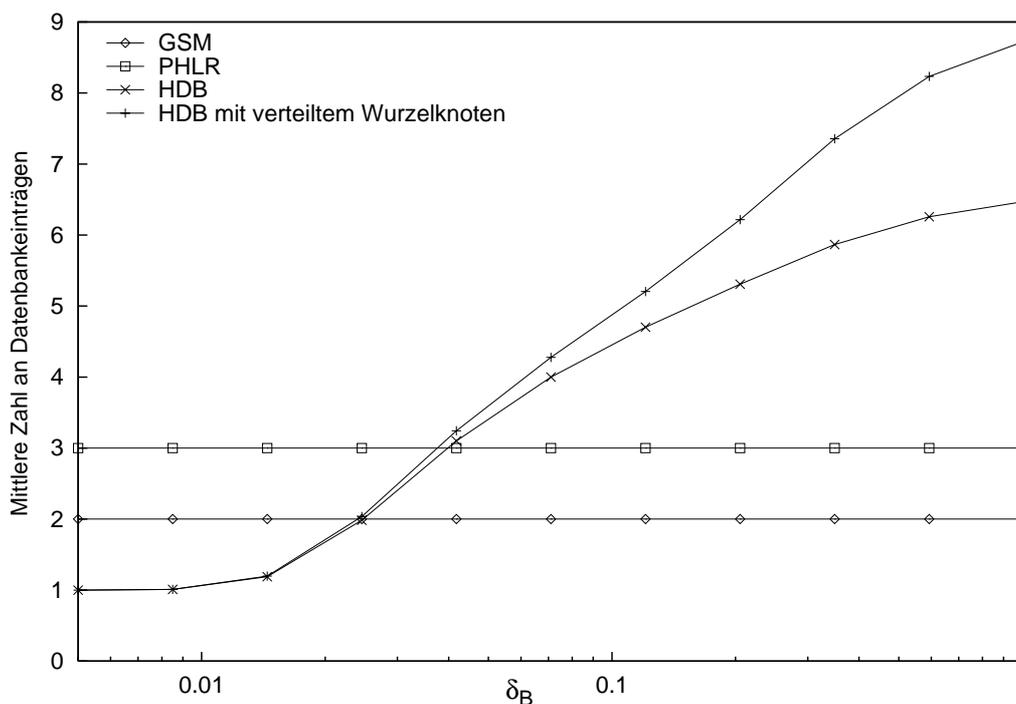


Abbildung 6-14: Vergleich der verschiedenen Verfahren zur Aufenthaltsortsdatenverwaltung bezüglich der mittleren Zahl an Datenbankeinträgen im System über dem Spreizfaktor δ_B der Bewegung für alle Teilnehmerklassen.

Abbildung 6-14 zeigt die mittlere Zahl an Datenbankeinträgen im System für GSM, für das PHLR-Verfahren und für das HDB-basierte Verfahren mit und ohne verteiltem Wurzelknoten. Bei GSM (\diamond) gibt es immer genau zwei Datenbankeinträge: einen im HLR und einen im VLR. Fügt man in diese zweistufige Hierarchie ein *Proxy HLR* (\square) als dritte Stufe ein, so ergeben sich immer genau drei Datenbankeinträge. Beim HDB-basierten Verfahren hängt die Zahl der Datenbankeinträge vom aktuellen Aufenthaltsort eines Teilnehmers ab. Gewichtet man diese Zahl mit der zugehörigen Zustandswahrscheinlichkeit $\pi_u^{(k)}$, so erhält man die dargestellten Werte. Beim Verfahren mit einem Wurzelknoten (\times) sind weniger Einträge als beim Verfahren mit verteiltem Wurzelknoten ($+$) notwendig, da im letzteren Fall Kopien der Einträge im Wurzelknoten eines Teilnehmers in den anderen Knoten der oberen Netzebene gehalten werden. Für kleine δ_B ergeben sich beim HDB-basierten Verfahren deutlich weniger Datenbankeinträge als bei GSM oder beim PHLR-Verfahren. Für sehr kleine δ_B ist beispielsweise meist nur ein Eintrag im Heimatknoten notwendig. Für mittlere und große δ_B dagegen ist eine zum Teil deutlich höhere Zahl notwendig. Für diesen Bereich wäre eine andere Form der Eintrags-

verzögerung – und damit auch ein verändertes Suchverfahren – sinnvoll, ohne daß dadurch der beobachtbare Verlauf und die Kosten für Suche und Ortsaktualisierung unterschiedlich würden. Anstelle der Zeigerkette vom Heimatknoten zum besuchten Knoten könnte man eine Zeigerkette vom Wurzelknoten zum besuchten Knoten verwenden. Beim Verfahren mit einem Wurzelknoten ergäben sich, je nach δ_B , Werte zwischen drei und vier für die Zahl der Datenbankseinträge. Beim verteilten Wurzelknoten ergäben sich jeweils genau sechs Einträge.

Die Abbildungen 6-15 und 6-16 zeigen, wie viele Datenbankzugriffe bei den verschiedenen Verfahren im Mittel je Ortsaktualisierung und je Teilnehmersuche erfolgen. Dabei werden lediglich die Zugriffe gezählt, die unmittelbar mit der Ortsdatenverwaltung zu tun haben. Die verbleibenden Zugriffe, die beim *Location Update* zur Teilnehmerauthentisierung und zum Transfer des Teilnehmerprofils zum aktuell besuchten VLR erfolgen, sind bei allen Verfahren gleich und auch nicht von δ_B oder δ_R abhängig. Sie werden deshalb nicht mitgezählt.

Bei GSM (\diamond) erfolgen je *Location Update* (siehe Abbildung 6-15) genau zwei Datenbankzugriffe: die Änderung des HLR-Eintrags und die Löschung im alten VLR. Beim PHLR-Verfahren (\square) gibt es aber immer wieder Fälle, bei denen vier Zugriffe je *Location Update* erfolgen, so daß im Mittel meist mehr als 2,5 Zugriffe je *Location Update* beobachtet werden können. Beim HDB-basierten Verfahren (\times bzw. $+$) sind beim Wechsel zwischen zwei Knoten in der Regel auch kaum mehr Zugriffe notwendig als beim PHLR-Verfahren.

Bei GSM (\diamond) erfolgen je Teilnehmersuche (siehe Abbildung 6-16) genau zwei Datenbankzugriffe, und zwar auf das HLR und auf das besuchte VLR zur Beschaffung der MSRN. Aus dem PHLR-Verfahren (\square) resultieren (mit $\delta_R = 0,05$) bereits für kleine δ_B relativ viele Fälle mit vier Zugriffen (auf das *Proxy HLR* in der Nähe des Ursprungs der Suche, auf das eigentliche HLR, auf das *Proxy HLR* in der Nähe des aktuellen Aufenthaltsorts des Teilnehmers und auf das besuchte VLR zur Beschaffung der MSRN), so daß hier im Mittel schon mehr als 2,5 Zugriffe je Teilnehmersuche erfolgen. Für große δ_B ergeben sich dann im Mittel sogar mehr als drei Zugriffe je Teilnehmersuche (siehe Abbildung 6-16). Dadurch, daß beim HDB-basierten Verfahren (\times bzw. $+$) bereits am Ursprung einer Suche ein Datenbankzugriff erfolgt, der selbst beim Ruftyp II und bei $\delta_R = 0,05$ nur in 20% der Fälle sofort zum Erfolg führt, beobachtet man hier deutlich mehr Datenbankzugriffe als beim PHLR-Verfahren. Führt man jedoch auch beim PHLR-Verfahren am Anfang einer Suche ein VLR-Lookup (\triangle) durch, so ergeben sich recht ähnliche Werte wie beim HDB-basierten Verfahren.

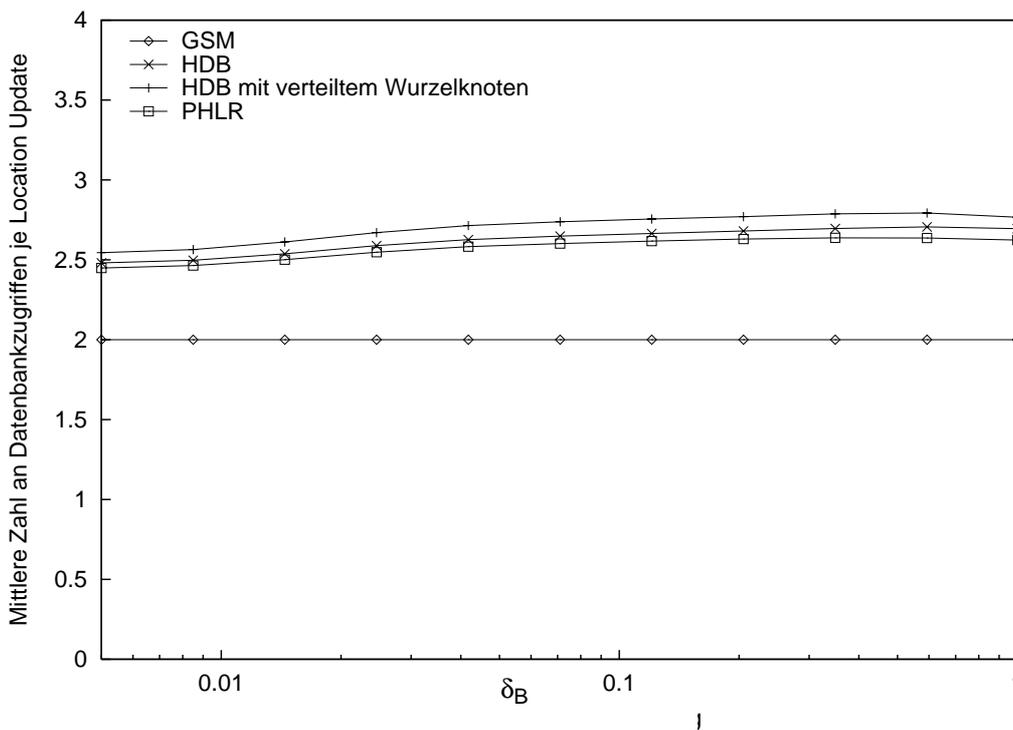


Abbildung 6-15: Vergleich der verschiedenen Verfahren zur Aufenthaltsortsdatenverwaltung bezüglich der mittleren Zahl an Datenbankzugriffen je *Location Update* über dem Spreizfaktor δ_B der Bewegung für alle Teilnehmerklassen.

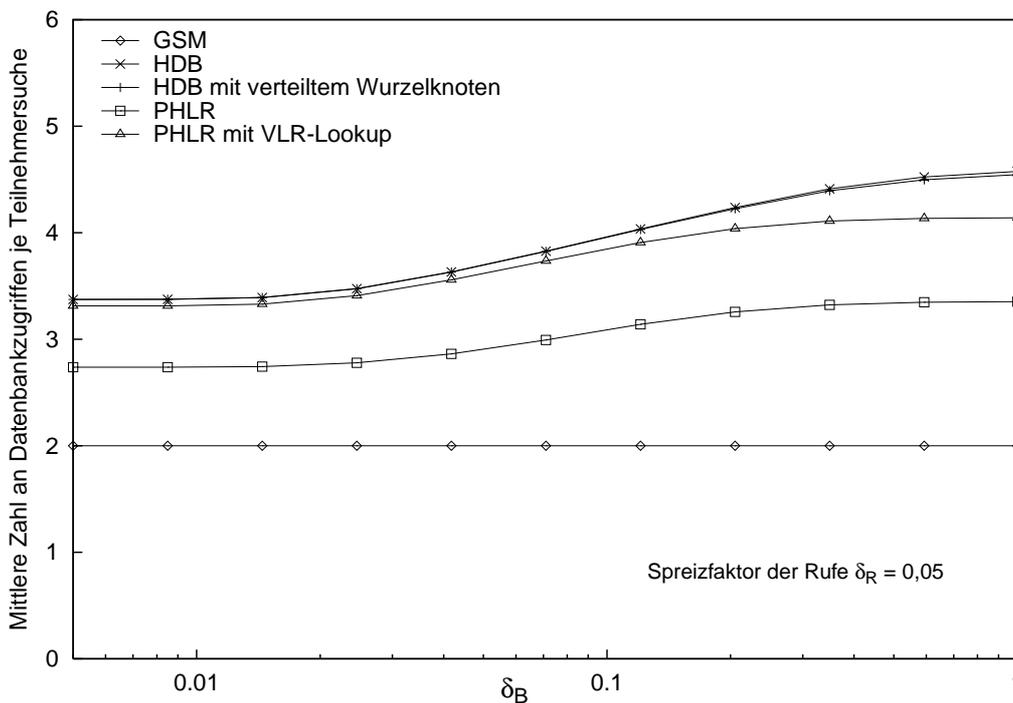


Abbildung 6-16: Vergleich der verschiedenen Verfahren zur Aufenthaltsortsdatenverwaltung bezüglich der mittleren Zahl an Datenbankzugriffen je Teilnehmersuche über dem Spreizfaktor δ_B der Bewegung für alle Teilnehmerklassen.

6.2.5 Verkehrsangebot je Netzinstanz

Bisher wurden die Ergebnisse global für das gesamte Netz betrachtet. Interessant ist natürlich auch, wie der durch die Mobilitätsverwaltung erzeugte Verkehr sich auf die verschiedenen Instanzen des Netzes verteilt. Diese Information ist außerdem eine wichtige Voraussetzung zur Berechnung von Antwortzeiten der verschiedenen Mobilitätsverwaltungsprozeduren. An dieser Stelle beschränken wir uns auf die Betrachtung des Verkehrsangebots, das in den Datenbankknoten verarbeitet werden muß. Ergebnisse bezüglich der Signalisierlast auf einzelnen Signalisierabschnitten sowie bezüglich der Bearbeitung der zugehörigen Signalisierprotokolle werden an dieser Stelle nicht dargestellt. Da sie Teil des Lastmodells sind, könnten sie natürlich genauso leicht ermittelt werden wie die nachfolgend präsentierten Ergebnisse.

Zur Vereinfachung der Darstellung wird zwischen Knoten der gleichen Ebene kein Unterschied gemacht. Durch die Berandung des Netzes ist zwar selbst bei Betrachtung aller Teilnehmerklassen das Verkehrsangebot an Knoten der gleichen Ebene nicht identisch, die Unterschiede sind aber durch das symmetrische Verhalten der Teilnehmer nicht so groß, daß sie eine getrennte Darstellung rechtfertigen würden. Folglich wird für die Darstellung das Verkehrsangebot aller Knoten einer Netzebene addiert und durch die Zahl der Knoten in der Netzebene dividiert. Die Tabellen 6-1 und 6-2 geben die zugehörigen Ergebnisse für zwei ausgewählte Punkte im Parameterraum wieder: für $\delta_B = 0,005$ und $\delta_R = 0,05$ sowie für $\delta_B = 1,0$ und $\delta_R = 0,1$.

Man sieht, daß für GSM immer dieselbe Last anfällt. Bei *Location Updates* sind die Knoten in der mittleren Ebene sowohl beim PHLR-Verfahren als auch beim HDB-basierten Verfahren identisch belastet und schirmen für den gesamten Bereich von δ_B die Knoten in der oberen Ebene recht gut von einem nahezu gleichbleibenden Teil der Last ab. Je nach Verfahren variiert die Zahl der Zugriffe je *Location Update* in Knoten der oberen Ebene leicht. So ergeben sich zum Beispiel bei verteiltem Wurzelknoten (WK) etwas höhere Werte als bei einem dedizierten Wurzelknoten je Teilnehmer. Bei höherem δ_B fällt in der mittleren und oberen Ebene etwas mehr Last durch *Location Updates* an.

Bei der Teilnehmersuche läßt sich zwischen Verfahren mit lokalem „Lookup“ und Verfahren ohne denselben unterscheiden. Bei $\delta_B = 0,005$ und $\delta_R = 0,05$ sind diese Verfahren (HDB, PHLR mit VLR-Lookup und GSM mit VLR-Lookup) in der Lage, einen wesentlichen Teil der Last durch Teilnehmersuchen von den Knoten der mittleren Ebene fernzuhalten (bzw. bei GSM mit VLR-Lookup, die Last in den HLRs zu reduzieren). Bei $\delta_B = 1,0$ und $\delta_R = 0,1$ dagegen ist der Erfolg nur gering oder verschwindet sogar völlig. Ähnlich verhält es sich mit der Fähigkeit der Knoten in der mittleren Ebene, Last von den Knoten der oberen Ebene fernzuhalten. Für $\delta_B = 0,005$ und $\delta_R = 0,05$ gelingt dies sehr gut. Für $\delta_B = 1,0$ und $\delta_R = 0,1$ jedoch erzeugen die Teilnehmersuchen beim HDB-basierten Verfahren sogar mehr Last in den Knoten der oberen Ebene als das GSM-Verfahren, und auch das PHLR-Verfahren reduziert die Last in den Knoten der oberen Ebene nur um knapp 20%.

$\delta_B = 0,005$ und $\delta_R = 0,05$	Zugriffe je Knoten je Ebene (ob/mitte/unt) bei 1024 <i>Location Updates</i>	Zugriffe je Knoten je Ebene (ob/mitte/unt) bei 1024 Teilnehmersuchen
GSM: 4 HLRs	512 / 0,00 / 4,0	256 / 0,00 / 1,0
dito – mit VLR-Lookup	(512 / 0,00 / 4,0)	202 / 0,00 / 1,8
PHLR	313 / 19,6 / 4,0	94 / 25,1 / 1,0
PHLR mit VLR-Lookup	(313 / 19,6 / 4,0)	94 / 18,5 / 1,8
HDB mit einem WK	321 / 19,6 / 4,0	110 / 18,5 / 1,8
HDB mit verteiltem WK	338 / 19,6 / 4,0	110 / 18,5 / 1,8

Tabelle 6-1: Mittlere Zahl an Datenbankzugriffen je 1024 Mobilitätsverwaltungsprozeduren für einen repräsentativen Netzknoten aus der oberen / mittleren / unteren Netzebene bei $\delta_B = 0,005$ und $\delta_R = 0,05$ sowie bei Überlagerung aller Teilnehmerklassen.

$\delta_B = 1,0$ und $\delta_R = 0,1$	Zugriffe je Knoten je Ebene (ob/mitt/unt) je 1024 <i>Location Updates</i>	Zugriffe je Knoten je Ebene (ob/mitt/unt) je 1024 Teilnehmersuchen
GSM: 4 HLRs	512 / 0,00 / 4,0	256 / 0,00 / 1,0
dito – mit VLR-Lookup	(512 / 0,00 / 4,0)	256 / 0,00 / 2,0
PHLR	336 / 21,0 / 4,0	210 / 29,1 / 1,0
PHLR mit VLR-Lookup	(336 / 21,0 / 4,0)	210 / 28,7 / 2,0
HDB mit einem WK	354 / 21,0 / 4,0	341 / 28,7 / 2,0
HDB mit verteiltem WK	372 / 21,0 / 4,0	323 / 28,7 / 2,0

Tabelle 6-2: Mittlere Zahl an Datenbankzugriffen je 1024 Mobilitätsverwaltungsprozeduren für einen repräsentativen Netzknoten aus der oberen / mittleren / unteren Netzebene bei $\delta_B = 1,0$ und $\delta_R = 0,1$ sowie bei Überlagerung aller Teilnehmerklassen.

6.3 Vergleich auf der Basis von Antwortzeiten

Wie bereits oben erwähnt, spielen bei der Leistungsbewertung eines Verfahrens neben der anfallenden Verkehrslast auch die Antwortzeiten der verschiedenen Verfahren zur Aufenthaltsortsdatenverwaltung eine wichtige Rolle. Dabei ist nicht unbedingt die Minimierung von Antwortzeiten notwendig. In der Regel reicht es aus, die Antwortzeiten unterhalb einer gewissen Grenze zu halten, ab der die Nutzer von Kommunikationsdiensten beginnen, die Antwortzeiten als ungewöhnlich wahrzunehmen. Letzteres gilt besonders für die Teilnehmersuche, da hierbei der sogenannte Rufverzug (*post-dialling delay*) als wichtiges Dienstgütekriterium empfunden wird. Bei der Aufenthaltsortsaktualisierung besteht die Gefahr, daß eine große Antwortzeit zu nicht-aktuellen Daten und damit zu einem erhöhten Aufwand bei zufälligerweise gleichzeitig erfolgenden Teilnehmersuchen führt.⁵

Anders als beim Vergleich der Verfahren im letzten Unterkapitel muß bei der Ermittlung der Antwortzeiten die gesamte Last aller Teilnehmer und aller Mobilitätsverwaltungsprozeduren berücksichtigt werden. Dazu ist es notwendig, für das untersuchte Beispiel sowohl die Zahl der Teilnehmer je Teilnehmerklasse, als auch die absoluten Raten von *Location Updates* und Teilnehmersuchen – was auch das Verhältnis zwischen den Raten beinhaltet – festzulegen bzw. zu ermitteln. Im folgenden gehen wir davon aus, daß je Teilnehmerklasse 100.000 Teilnehmer existieren (d. h. daß das Netz eine Population von ca. 100 Millionen hat), daß jeder Teilnehmer 0,0001 Rufe pro Sekunde (0,36 Rufe/h) erhält und daß er δ_B *Location Updates* je kommendem Ruf durchführt.⁶ Letzteres bedeutet, daß der Kehrwert der sogenannten *Call-to-Mobility Ratio* (CMR) im untersuchten Beispiel mit dem Spreizfaktor δ_B der Bewegung korrespondiert und damit Werte zwischen 0,005 und 1,0 annimmt. Darüberhinaus ist es notwendig – beispielsweise bei der Untersuchung von Verzögerungen im Signalisierernetz – auch weiteren Nachrichtenverkehr, wie etwa Signalisierung zur Rufsteuerung, zu berücksichtigen.

6.3.1 Methodik zur Bestimmung von Antwortzeiten

Prinzipiell lassen sich Antwortzeiten mit simulativen oder analytischen Verfahren bestimmen. Das vorgestellte Modell zur Beschreibung von Teilnehmerverhalten, Verkehrslast und Netzverhalten ist sehr gut für simulative Studien geeignet. Bei sehr großen Systemen und beim Untersuchen eines breiten Bereichs des Parameterraums sind analytische Verfahren in der Regel jedoch sehr viel schneller und bieten einen Überblick über das generelle Verhalten. Andererseits sind bei komplexen Systemen exakte und umfassende Ergebnisse nicht oder nur mit sehr großem Aufwand ermittelbar.

Als Beispiel für ein analytisches Verfahren wird hier zur Leistungsbewertung ein einfaches Dekompositions- und Aggregationsverfahren zur Bestimmung von Mittelwerten eingesetzt. Dekompositionsverfahren [Cou77, Kü76, Kü77, Whi83] zerlegen ein System in Teilsysteme und untersuchen diese getrennt voneinander. Die gegenseitige Beeinflussung der Teilsysteme wird dabei in der Regel vereinfacht. Dies geschieht, indem beispielsweise der Ankunftsprozeß an einem Teilsystem als ein Erneuerungsprozeß angenommen wird, dessen Parameter durch Aggregation von Ausgangsprozessen anderer Teilsysteme bestimmt werden. Dabei werden meist nur wenige charakteristische Größen, wie zum Beispiel die ersten beiden Momente, berücksichtigt.

Für das hier untersuchte Beispiel wird nun angenommen, daß der Ankunftsprozeß an einer verarbeitenden Instanz durch einen Poisson-Prozeß angenähert werden kann. Diese Annahme erleichtert die Analyse der Durchlaufdauer durch eine verarbeitende Instanz in hohem Maße. Die Analyse von Signalisiererverkehr in SS7-basierten Signalisierernetzen wurde unter der selben

-
5. Unter Umständen kann die Suche nach einem Teilnehmer sogar auf Grund von noch nicht aktualisierten Ortsdaten scheitern.
 6. Hier wird von einem stationären Verkehrsangebot ausgegangen, das beispielsweise in einer Hauptverkehrsstunde angetroffen werden könnte.

Annahme bereits erfolgreich durchgeführt [BKW94, BKW93, Baf95, KS96]. Die Annahme wird für das hier untersuchte System durch eine Reihe von Beobachtungen gerechtfertigt:

- Das Verkehrsangebot im untersuchten System wird durch Überlagerung einer Vielzahl unabhängiger Quellen (100 Millionen Teilnehmer, die durch unabhängige *Location-Update*- und Teilnehmersuchprozesse modelliert werden) gebildet.⁷
- Für jeden Netzknoten gilt, daß er von einer Vielzahl von Eingangsprozessen gespeist wird. Ein Knoten der mittleren Ebene erfährt beispielsweise eine Überlagerung von Verkehren, die von 16 untergeordneten Knoten, vier Nachbarknoten und einem übergeordneten Knoten stammen.
- Bearbeitende Knoten besitzen in der Regel mehrere Prozessoren und im Netzmodell wird die Verteilung von Last durch eine zufällige Verteilung von Anforderungen auf Prozessoren modelliert.
- Der Ausgangsprozeß eines Knotens verteilt sich meist auf mehrere Ziele. Bei einem Knoten der mittleren Ebene sind dies wiederum alle Nachbarn nach oben, unten und zur Seite. Bei den Knoten der unteren Ebene verteilt sich der Verkehr auf den übergeordneten Knoten und auf eine Nachrichtensenke.

Nimmt man nun noch die Annahme hinzu, daß die Durchlaufdauern einer Anforderung durch verschiedene Teilsysteme unabhängig sind, so lassen sich die Ende-zu-Ende-Durchlaufdauern bzw. die Antwortzeiten der Mobilitätsverwaltungsprozeduren durch Addition der unabhängigen Zufallsvariablen der einzelnen Durchlaufdauern ermitteln. Die Abbildung von *Physical Entity Actions* auf Prozessoren in Kapitel 5.4.2 wurde explizit so definiert, daß diese Vorgehensweise möglich bleibt.

Bei der Analyse eines einzelnen Teilsystems ist die Untersuchung relativ komplexer Modelle möglich. Das in Kapitel 5.4.2 vorgestellte Modell eines Netzknotens wurde so erstellt, daß zur Ermittlung von mittleren Durchlaufdauern die sogenannte Momentenmethode für ein $\Sigma M_i/GI_i/1$ -System mit Prioritäten und Rückkopplungen verwendet werden kann, wie es in [Pat90] beschrieben ist. Die Momentenmethode beruht im wesentlichen auf einer geschickten Kombination des Gesetzes von Little [Lit63], des PASTA-Theorems [Wol82], des Ergebnisses für die Vorwärtsrekurrenzzeit aus der Erneuerungstheorie und der Bestimmung eines mittleren Systemzustands bei Rückkopplung. In [Bod94] wurde eine numerische Vorgehensweise entwickelt, mit der auch sehr komplexe derartige Modelle sehr schnell analysiert werden können und im Rahmen eines Werkzeugs zur Leistungsbewertung von „Intelligenten Netzen“ eingesetzt [Kan94, Pol96]. Darüberhinaus wurde das Verfahren um die Behandlung gesättigter Systeme erweitert [Bod94, Bod96] und auch bereits auf die Untersuchung von SS7-Verkehr angewendet [BS97].

7. An dieser Stelle sei nochmals darauf hingewiesen, daß zur Modellierung des korrelierten Verhaltens einiger Teilnehmer eine Gruppenbildung, wie in Kapitel 5.1.5.1 vorgeschlagen, vorgenommen werden kann. Bei der analytischen Behandlung kann dies vereinfacht durch Gruppenankünfte modelliert werden. Die Poisson-Eigenschaft des Ankunftsprozesses wird dadurch nicht beeinträchtigt. Lediglich die Gruppengröße bedarf gegebenenfalls einer Anpassung an das untersuchte Teilsystem.

Die auf diese Weise erzielten Ergebnisse können in jedem Fall dazu dienen, die qualitativen Unterschiede zwischen den untersuchten Verfahren zur Aufenthaltsortsdatenverwaltung zu ermitteln. Die aufgeführten Beobachtungen lassen jedoch darüberhinaus darauf schließen, daß die gemachten Annahmen vertretbar sind und daß folglich die erzielten Ergebnisse sehr nahe an den tatsächlich zu erwartenden Werten liegen.

6.3.2 Parameter des Netzmodells und Bestimmung von Antwortzeiten

Es wird ein sehr einfaches Netzmodell mit willkürlich gewählten Parametern verwendet. Der Transport von Signalisiermeldungen und die Bearbeitung der zugehörigen Signalisierprotokolle wird durch eine konstante Verzögerung von 5 ms je Signalisierabschnitt modelliert. Damit ergibt sich der Teil der Antwortzeit, der von Kommunikation zwischen Netzknoten herrührt, aus der Multiplikation von 10 ms mit der *Zahl der „Hops“*. Die Berücksichtigung von zusätzlichem Signalisierverkehr ist hierbei nicht notwendig. Bei der Verwendung komplexerer Signalisiernetzmodelle müßte dies jedoch geschehen und wäre auch sehr einfach möglich. Die zugehörigen Berechnungen könnten beispielsweise mit Hilfe des in [Baf95, Kan94] entwickelten Werkzeugs zur Leistungsbewertung von SS7-Netzen durchgeführt werden.

Für die Bearbeitung von lesenden und schreibenden Operationen in Datenbankknoten wird eine konstante (D) Bearbeitungszeit von $h = h_{\text{Suche}} = h_{\text{LU}s} = 20$ ms angenommen. Für alle Szenarien wird gleichermaßen angenommen, daß in den Knoten auf der unteren Ebene diese Operationen in je 2 Prozessoren, in den Knoten der mittleren Ebene in je 10 Prozessoren und in den Knoten der oberen Ebene in je 200 Prozessoren bearbeitet werden. Die Operationen werden jeweils zufällig auf die Prozessoren verteilt. Auch hier wären sehr viel komplexere Modelle denkbar, die beispielsweise die Bearbeitungszeit von der Größe des Datenbestands abhängig machen, unterschiedlichen Operationen unterschiedliche Bearbeitungszeiten zuweisen oder unterschiedliche Operationen von unterschiedlichen Prozessoren bearbeiten lassen.

Wie schon im letzten Unterkapitel wird das Verkehrsangebot an einen Datenbankknoten aus dem Mittel der Angebote an allen Knoten der gleichen Netzebene berechnet. Zwar ist durch die Berandung des Netzes das Verkehrsangebot an Knoten der gleichen Ebene nicht exakt gleich groß, für die eher qualitative Betrachtung der Antwortzeiten braucht dieser Effekt jedoch nicht berücksichtigt zu werden. Die mittlere Antwortzeit einer Operation an einem Knoten (s_{LU} bzw. s_{Suche}) ergibt sich damit unmittelbar aus dem Verkehrsangebot ρ_{Suche} , das von Teilnehmersuchen an einem Prozessor herrührt, und dem Verkehrsangebot $\rho_{\text{LU}s}$, das von *Location Updates* an einem Prozessor herrührt:

$$s_{\text{Suche}} = s_{\text{LU}s} = h + \frac{RBZ}{1 - \rho_{\text{Suche}} - \rho_{\text{LU}s}} \quad (6-1),$$

wobei *RBZ* die mittlere Restbedienzeit eines Prozessors ist:

$$RBZ = \frac{h}{2} \cdot (\rho_{\text{Suche}} + \rho_{\text{LU}s}) \quad (6-2).$$

Priorisiert man Teilnehmersuchen vor *Location Updates* nicht-unterbrechend, so ergibt sich:

$$s_{\text{Suche}} = h + \frac{RBZ}{1 - \rho_{\text{Suche}}} \quad (6-3),$$

$$s_{\text{LU}s} = h + \frac{RBZ}{(1 - \rho_{\text{Suche}} - \rho_{\text{LU}s}) \cdot (1 - \rho_{\text{Suche}})} \quad (6-4).$$

6.3.3 Ergebnisse

Die Abbildungen 6-17 und 6-18 zeigen die mittleren Antwortzeiten für ausgewählte Verfahren zur Aufenthaltsortsdatenverwaltung bei gleicher Priorität für *Location Updates* und Teilnehmersuchen und für $\delta_R = 0,05$.

Die mittleren Antwortzeiten für *Location Updates* (siehe Abbildungen 6-17) bei „optimiertem“ GSM mit den HLRs in der mittleren Netzebene (\triangle), beim *Proxy-HLR*-Verfahren (\square), beim „optimierten“ *Proxy-HLR*-Verfahren ($+$) und beim HDB-basierten Verfahren (\times) sind über den gesamten untersuchten Parameterbereich jeweils alle ungefähr gleich groß. Für große δ_B werden die mittleren Antwortzeiten sehr groß. Dies läßt sich damit erklären, daß die Knoten in der mittleren Netzebene nicht leistungsfähig genug ausgelegt sind. Die Antwortzeiten beim GSM-Verfahren mit den HLRs in der oberen Netzebene (\diamond) sind für kleinere δ_B größer als bei den anderen Verfahren. Für große δ_B dagegen steigen sie nicht auf so hohe Werte an. Dies liegt daran, daß hierbei keine Knoten der mittleren Ebene beteiligt sind, die als Engpaß wirken könnten, und daß die Knoten der unteren und oberen Ebene ausreichend dimensioniert sind.

Bezüglich der mittleren Antwortzeiten von Teilnehmersuchen (siehe Abbildung 6-18) unterscheiden sich die verschiedenen Verfahren deutlicher. Die kürzesten Antwortzeiten haben die GSM-Verfahren (\triangle und \diamond). Die *Proxy-HLR*-Varianten ($+$ und \times) haben im Vergleich dazu deutlich höhere mittlere Antwortzeiten. Das HDB-basierte Verfahren schließlich hat die längsten mittleren Antwortzeiten. Auch bei den Antwortzeiten von Teilnehmersuchen macht sich der Effekt der hoch belasteten Datenbankknoten in der mittleren Netzebene bei großem δ_B bemerkbar. Es wäre sinnvoll, in diesem Bereich die Antwortzeit der Teilnehmersuche zu reduzieren, da sie als Teil des Rufverzugs Einfluß auf die Dienstgüte hat. Am effektivsten wäre die Erhöhung der Leistungsfähigkeit der Datenbankknoten in der mittleren Netzebene. Alternativ kann man jedoch auch die Bearbeitung von Teilnehmersuchen vor der Bearbeitung von *Location Updates* priorisieren. Dadurch werden zwar die Antwortzeiten von *Location Updates* weiter erhöht, die für die Dienstgüte wichtigeren Antwortzeiten von Teilnehmersuchen werden jedoch verringert..

Die Abbildungen 6-19 und 6-20 zeigen die Ergebnisse für die mittleren Antwortzeiten, wenn man die Datenbankoperationen zur Teilnehmersuche nicht-unterbrechend vor den Operationen zur Datenaktualisierung priorisiert. Qualitativ ergeben sich ähnliche Ergebnisse wie in den

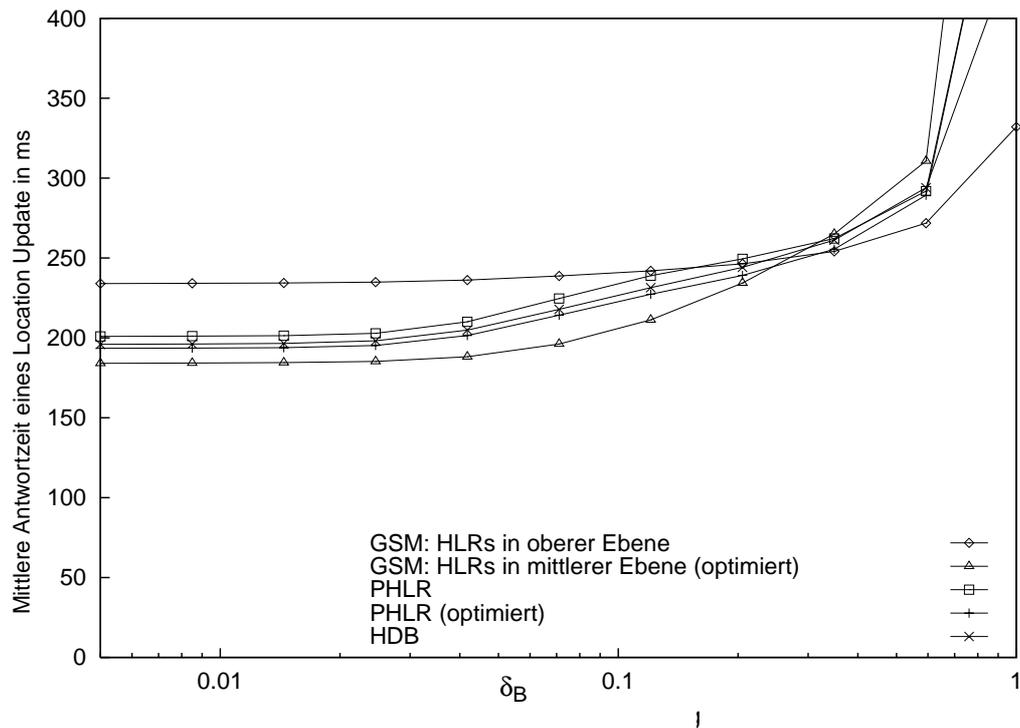


Abbildung 6-17: Vergleich der Antwortzeiten von *Location Updates* für verschiedene Verfahren zur Aufenthaltsortsdatenverwaltung.

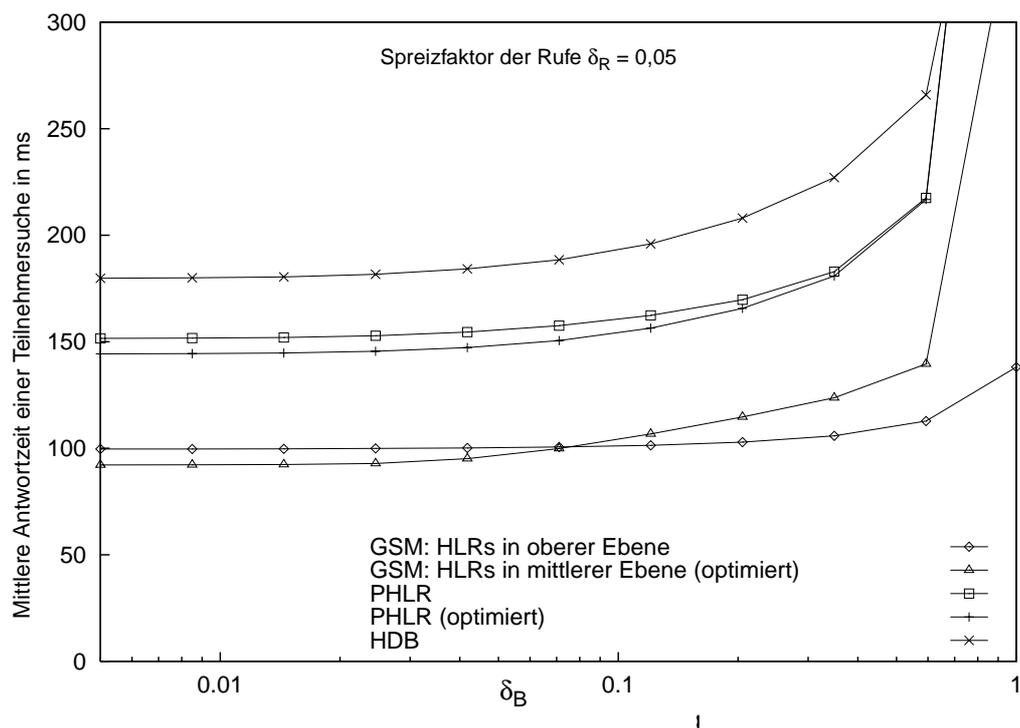


Abbildung 6-18: Vergleich der Antwortzeiten von Teilnehmersuchen für verschiedene Verfahren zur Aufenthaltsortsdatenverwaltung.

Abbildungen 6-17 und 6-18 mit dem Unterschied, daß bei der Teilnehmersuche auch für große Mobilität der Teilnehmer – d.h. für große δ_B und eine entsprechende *Call-to-Mobility Ratio* – die Antwortzeiten von Teilnehmersuchen relativ klein bleiben

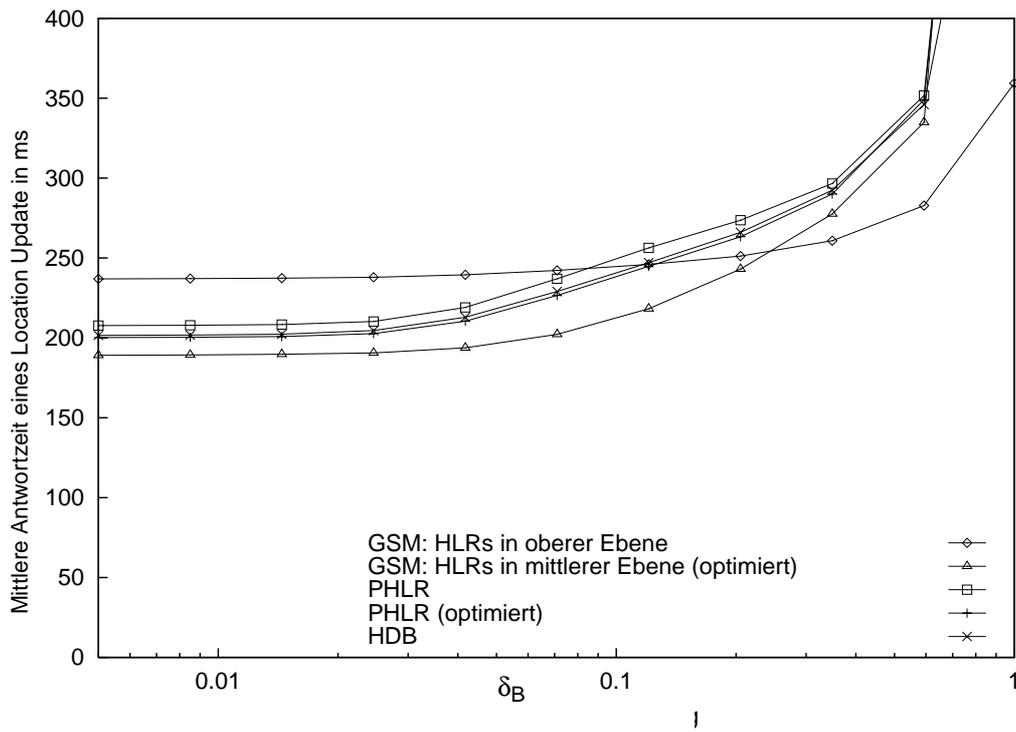


Abbildung 6-19: Vergleich der Antwortzeiten von *Location Updates* für verschiedene Verfahren zur Aufenthaltsortsdatenverwaltung bei Priorisierung von Teilnehmersuchoperationen.

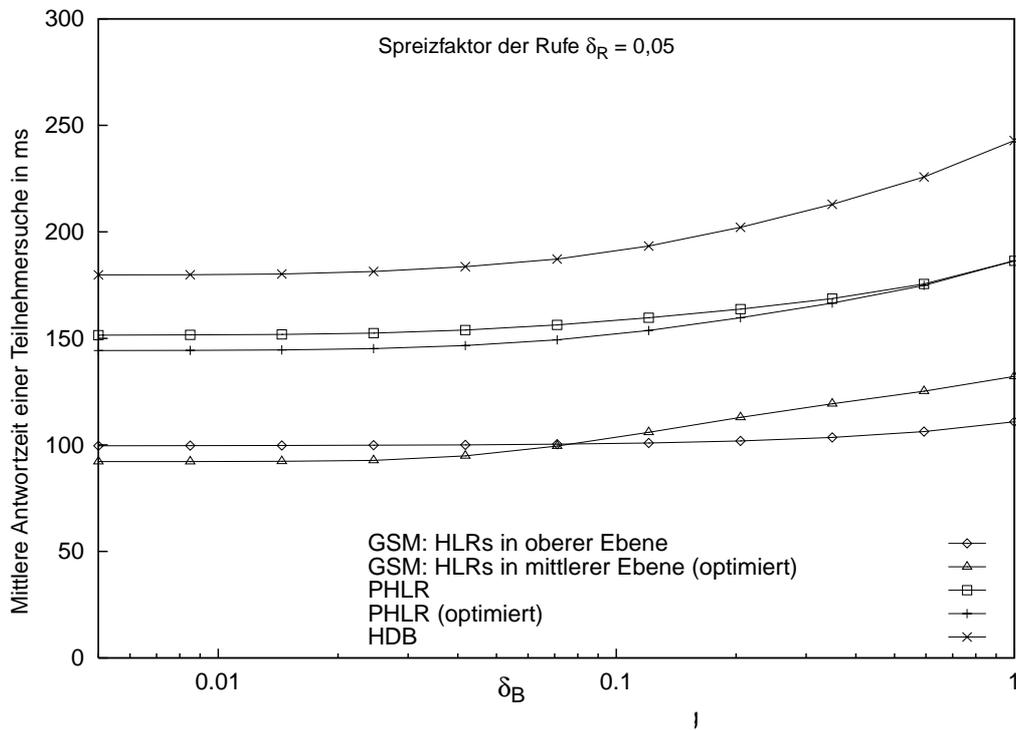


Abbildung 6-20: Vergleich der Antwortzeiten von Teilnehmersuchen für verschiedene Verfahren zur Aufenthaltsortsdatenverwaltung bei Priorisierung von Teilnehmersuchoperationen.

6.4 Fazit

In diesem Kapitel wurde exemplarisch gezeigt, wie das in der Arbeit entwickelte Teilnehmermodell verwendet werden kann, um verschiedene Verfahren zur Ortsdatenverwaltung zu bewerten. Dazu wurde beispielhaft eine Netzstruktur angenommen, Parameter des Netzmodells willkürlich gewählt und die Parameter des Teilnehmermodells variiert. Die damit erzielten Ergebnisse illustrieren, wie erhöhter Aufwand bei der Bearbeitung von Mobilitätsverwaltungsprozeduren und erhöhter Ressourcenbedarf bei der Datenhaltung zu einer Reduzierung von Signalerlast führen kann. Abhängig davon, wie groß dieser Aufwand ist und wie leistungsfähig Netzknoten ausgelegt werden können, kann dieser erhöhte Aufwand jedoch auch in erhöhten Antwortzeiten von Mobilitätsverwaltungsprozeduren resultieren.

Gleichzeitig zeigt das untersuchte Beispiel, daß aufwendigere Verfahren nicht unbedingt zu signifikanten Einsparungen bei der Signalisierlast führen müssen. Die Effektivität eines lokalen VLR-Lookup vor dem Auslösen einer Teilnehmersuche beispielsweise hängt stark von den Parametern des Teilnehmermodells ab. Auch eine große Zahl von Hierarchiestufen bei der Aufenthaltsortsdatenverwaltung ist nicht unbedingt besonders effizient. So schneidet das HDB-basierte Verfahren für die untersuchte Netzstruktur und die angenommene Datenbankhierarchie nicht wesentlich besser ab als das *Proxy-HLR*-Verfahren, das eine geringere Zahl von Hierarchiestufen hat. Prinzipiell sind sich die beiden Verfahren recht ähnlich. Würde man die Zahl der Hierarchiestufen beim HDB-basierten Verfahren reduzieren, so ergäben sich – abgesehen von der mittleren Zahl an Datenbankeinträgen – prinzipiell identische Ergebnisse wie beim *Proxy-HLR*-Verfahren.

Für eine willkürliche Zuordnung von Teilnehmern zu HLRs ergeben sich für das *Proxy-HLR*-Verfahren ähnlich gute Ergebnisse wie bei einer optimierten Zuordnung. Das vorgestellte HDB-basierte Verfahren dagegen beruht auf dem Konzept eines Heimatknotens, was eine freie Zuordnung von Kennzeichnern und Nummer (bzw. „Wunschrufnummern“) zu Teilnehmern erschwert. Verzichtet man auf diese freie Zuordnung, so kann das klassische GSM-Verfahren mit einer optimierten Zuordnung von Teilnehmern zu HLRs und eine günstige Platzierung dieser HLRs über weite Teile des untersuchten Parameterraums sehr gute Ergebnisse erzielen, und zwar sowohl bezüglich des Signalisieraufkommens als auch bezüglich der Antwortzeiten. Die deterministische Zahl von Datenbankzugriffen sorgt beim GSM-Verfahren darüberhinaus für eine geringere Variabilität der Antwortzeiten. Bei den anderen Verfahren dagegen hängt die Zahl der Datenbankzugriffe von den Knoten ab zwischen denen die Mobilitätsverwaltungsprozeduren ablaufen.

Kapitel 7

Zusammenfassung und Ausblick

Informations- und Kommunikationsdienste werden in Zukunft in hohem Maße personenbezogen sein. Teilnehmern werden diese Dienste zu jeder Zeit und an jedem Ort zur Verfügung gestellt werden. Dabei stellt die Unterstützung der Mobilität von Personen und Endgeräten in einem heterogenen Umfeld einen wesentlichen Baustein zur Dienstleistung dar. Um netze- und netztechnikenübergreifende Mobilität für eine große Zahl von Personen zu ermöglichen, bedarf es leistungsfähiger und skalierbarer Verfahren zur Mobilitätsverwaltung. In der vorliegenden Arbeit werden das Phänomen Mobilität im Rahmen des OSI-Referenzmodells dargestellt und die grundlegenden Mechanismen und Möglichkeiten zur Unterstützung von Mobilität sowohl in klassischen Telekommunikationsnetzen als auch in IP-basierten Netzen systematisiert. Das Spannungsfeld zwischen Wegführung, Zielsuche und Datenaktualisierung wird dargestellt und basierend auf diesen Überlegungen wird eine Reihe von Vorschlägen zur Weiterentwicklung bestehender Mobilitätsverwaltungsverfahren erarbeitet. Drei Vorschläge sind dabei besonders hervorzuheben: Das Senden von frühen *Binding Updates* in *Mobile IP*, die Verwendung von *Proxy HLRs* in GSM/UMTS sowie die Einrichtung einer *Common Distributed Database* für netzeübergreifende Mobilitätsverwaltung.

Zum quantitativen und qualitativen Vergleich verschiedener Verfahren zur Mobilitätsverwaltung bedarf es einer problemangepaßten Modellierung des Verhaltens der mobilen Instanzen. In der Arbeit wird ein entsprechendes Teilnehmermodell vorgeschlagen, das sowohl die Bewegung von Personen als auch die Suche derselben – und damit „kommende Rufe“ – modelliert. Einen besonders wichtigen Beitrag leistet die Arbeit dabei durch das Entwickeln leistungsfähiger Verfahren zur Parametrisierung des Bewegungsmodells sowie durch die Einführung verschiedener Rufstypen bei der Modellierung der Teilnehmersuche, die den Ursprung von Rufen in Abhängigkeit vom Bewegungsverhalten der Teilnehmer beschreiben. In Verbindung mit einem geeigneten Lastmodell wird das Teilnehmermodell dazu benutzt, Modellverkehr zu erzeugen, der zum Vergleich verschiedener Mobilitätsverwaltungsverfahren verwendet werden kann. Der Transport von Steuermeldungen und die Ausführung der Mobilitätsverwaltungsprozeduren in einem verteilten System wird mittels eines flexiblen Netzmodells dargestellt. Dieses Netzmodell ist so ausgelegt, daß eine analytische Leistungsbewertung, die auf einem komplexen Prozessmodell beruhen kann, ermöglicht wird.

Am Ende der Arbeit wird die GSM-Mobilitätsverwaltung mit Verfahren, die im Rahmen von GSM oder UMTS alternativ oder als Fortentwicklung Einsatz finden können, verglichen. Dabei wird einerseits der erzeugte Modellverkehr in Abhängigkeit von den Parametern des Teilnehmermodells als Kriterium herangezogen und andererseits die zu erwartende Antwortzeit verschiedener Prozeduren als einzuhaltende Randbedingung berücksichtigt.

Sowohl bei der Modellbildung, als auch bei der abschließenden Leistungsbewertung wird in der Arbeit der untersuchte Parameterraum bewußt groß gehalten. Die Ergebnisse liefern dadurch Aussagen über die Leistungsfähigkeit der untersuchten Mobilitätsverwaltungsverfahren für ein breites Spektrum von Einsatzmöglichkeiten. Mobile Dokumente im *World Wide Web*, mobile Agenten oder mobile *computational objects* zeigen in der Regel ein anderes Verhalten als Teilnehmer in einem funkbasierten Mobilkommunikationsnetz oder mobile Rechner im Internet. Diese zeigen wiederum ein anderes Verhalten als UPT-Teilnehmer, als das Ziel einer *Freephone*-Nummer (0800) oder als das Ziel einer netzbetreiberunabhängigen Rufnummer. Manche Verfahren eignen sich nur für einen kleinen Teil des Parameterraums, können dort jedoch deutliche Vorteile vor anderen Verfahren aufweisen. Untersuchungen für spezifische Anwendungen, für die der Parameterraum auf bestimmte Bereiche eingeschränkt werden kann, wurden in der Arbeit nur in sehr begrenztem Maße durchgeführt. Die Anwendung der Methodik für spezielle Systeme würde auch hierfür entsprechende Aussagen über die Leistungsfähigkeit verschiedener Verfahren liefern und könnte damit eine Entscheidung über ihren Einsatz unterstützen.

Die Arbeit beschränkt sich bei der Leistungsuntersuchung auf einige wenige Verfahren, die auch analytisch relativ einfach zu behandeln sind. Der Einsatz von Simulationen und von anspruchsvolleren Analysetechniken wird durch die generische Modellbildung aber ebensogut ermöglicht. Damit könnten weitergehende Aussagen über die Leistungsfähigkeit von Verfahren getroffen werden. So könnten beispielsweise für die Antwortzeiten auch die zugehörigen Verteilungen ermittelt werden, und es könnten Verfahren untersucht werden, bei denen der Verlauf der Mobilitätsverwaltungsprozeduren in stärkerem Maße von der Vergangenheit der Bewegung oder von der Korrelation mit dem Verhalten anderer Teilnehmer bestimmt ist.

Wie die Diskussion der Ergebnisse der Leistungsuntersuchung gezeigt hat, sind neben der rein netztechnischen Leistungsfähigkeit eines Verfahrens beispielsweise auch die Verwaltbarkeit, die Flexibilität und die Kompatibilität mit existierenden Verfahren wichtige Kriterien zur Bewertung. So kann etwa aus den Ergebnissen in Kapitel 6 geschlossen werden, daß der Einsatz von *Proxy HLRs* der Einführung von neuen hierarchischen Datenbankstrukturen in vielen Fällen vorgezogen werden kann, da die im Vergleich relativ geringen Einbußen bei der Leistungsfähigkeit durch höhere Flexibilität bei der Zuordnung von Rufnummern zu Teilnehmern und durch Kompatibilität mit bestehenden GSM-Protokollen aufgewogen werden. Für eine weitergehende Diskussion solcher Kriterien – zu denen beispielsweise auch der Schutz von personenbezogenen Daten gehört – bietet der erste Teil der Arbeit zwar eine gute technische und systematische Grundlage, kann aber diesbezüglich keinen erschöpfenden Beitrag leisten.

Abschließend seien die in der Einleitung bereits aufgeführten Hauptcharakteristika der Arbeit noch einmal kurz zusammengefaßt:

- ❑ Die Arbeit befaßt sich nicht mit Mobilität innerhalb von Funkzellen oder zwischen Funkzellen sondern mit dem Teil der Mobilität, der über die Begrenzungen eines einzelnen Zugangssystems hinausgeht, bei dem sich in der Regel die Adresse eines Endsystems in der Vermittlungsschicht ändert.
- ❑ Bei dieser Form der Mobilität spielt die verteilte Natur des betrachteten Systems eine wichtige Rolle. Dies gilt insbesondere für die verteilte Datenhaltung und die Verteilung von Ursprung und Ziel von Mobilitätsverwaltungsprozeduren.
- ❑ Bisherige Veröffentlichungen beschränken sich in der Regel auf einen einzelnen Vorschlag zur Mobilitätsverwaltung, ohne einen Überblick über Gemeinsamkeiten und Unterschiede zwischen verschiedenen Netzen und Netztechniken zu geben. Die vorliegende Arbeit beschreibt daher zuerst, welche grundlegenden Mechanismen zur Unterstützung von Mobilität in Kommunikationsnetzen vorhanden sein müssen, und systematisiert dann die verschiedenen Optionen, die sich beim Entwurf der entsprechenden Verfahren bieten. Dies geschieht netztechnikenübergreifend und kann deshalb als eine wesentliche Grundlage für den Entwurf von Mobilitätsverwaltungskonzepten in zukünftigen heterogenen Netzstrukturen dienen.
- ❑ Die Tragfähigkeit der vorgestellten Systematisierung wird durch eine Reihe von neuen, leistungsfähigen Vorschlägen zur Mobilitätsverwaltung untermauert.
- ❑ Es wird ein Vorschlag für ein allgemein einsetzbares, problemangepaßtes Teilnehmermodell gemacht, das alle wesentlichen Aspekte des Teilnehmerverhaltens berücksichtigt. Es kann daher als Referenzmodell zum Vergleich verschiedener Mobilitätsverwaltungsverfahren eingesetzt werden. (Der Einsatz als Referenzmodell ist hierbei nicht auf Telekommunikationsteilnehmer beschränkt. Bei entsprechender Wahl der Parameter kann beispielsweise auch das Verhalten mobiler Objekte im Internet oder der Verkehr für eine IN-Rufnummer beschrieben werden).
- ❑ Es wurden schnelle und leistungsfähige Verfahren zur Parametrisierung des Teilnehmermodells entwickelt und beschrieben. Besonders hervorzuheben sind hierbei die vorgeschlagenen Verfahren zur schnellen Parametrisierung des Bewegungsmodells. Dies ist besonders wichtig, da nur so große Netze und Systeme mit vertretbarem Aufwand untersucht werden können.
- ❑ Das Teilnehmermodell wird durch ein Lastmodell und ein flexibles Netzmodell ergänzt. Damit wurde für die Leistungsbewertung verschiedener Verfahren ein geeigneter Rahmen geschaffen.
- ❑ Bei einer exemplarisch durchgeführten Leistungsbewertung für ein Netz mit ca. 100 Millionen Teilnehmern und 1024 Orts- bzw. Mobilfunkvermittlungsstellen werden wichtige Erkenntnisse über den Einsatz fortentwickelter Verfahren im Rahmen der GSM- und der UMTS-Mobilitätsverwaltung gewonnen.

Literaturverzeichnis

- [AH95] Akyildiz, I. F., Ho, J. S. M.: *Dynamic mobile user location update for wireless PCS networks*, Wireless Networks, Vol. 1, No. 2, ACM-Press, Baltzer Science Publishers, July 1995, pp. 187–196.
- [AHL96] Akyildiz, I. F., Ho, J. S. M., Lin, Y. B.: *Movement-Based Location Update and Selective Paging for PCS Networks*, IEEE/ACM Transactions on Networking, Vol. 4, No. 4, August 1996, pp. 629–638.
- [AMM96] Adamidis, E. P., Manolesos, Y. I., Markoulidakis, J. G., Theologou, M. E., Sykas, E. D.: *Mobility Control Plane for COBUCCO*, Proceedings of ACTS Mobile Telecommunications Summit: Volume 2, Granada, Spain, November 1996, pp. 835–841.
- [Ape83] Apers, P. M.: *Query processing and data allocation in distributed database systems*, Mathematical Centre Tracts 156, Mathematisch Centrum, Amsterdam, 1983.
- [ATMF97] ATM-Forum 97-1041: *Reference Configuration, Requirements and an Architectural Model for Location Management in WATM Networks*, ATM Forum, December 1997.
- [Baf95] Bafutto, M.: *Modellierung, Verkehrsanalyse und Planung von Zentralkanal-Signaliernetzen im ISDN mit besonderer Berücksichtigung neuer Dienste der Mobilkommunikation und des Intelligenten Netzes*, Dissertationsschrift, 63. Bericht über verkehrstheoretische Arbeiten, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, 1995.
- [BDW97] Bostelmann, G., Depouilly, B., Weber, A.: *Netzarchitekturen für Multimediamobilität*, Alcatel Telecom Rundschau, Nummer 3, 1997, S. 222–231.
- [BG92] Bell, D. A., Grimson, J. B.: *Distributed database systems*, Addison-Wesley, Wokingham, England, 1992.
- [BG98] Bhat, R. R., Gupta, R., *ATM-Forum 98-0402: Baseline document for Wireless ATM CSI*, ATM Forum, July 1998.
- [BGH90] Bowen, T., Gopal, G., Herman, G.: *A Scalable Database Architecture for Network Services*, Proceedings of the ISS'90, 1990, paper B7#3, Vol. V, pp. 45–51.
- [BHJK96] Berger, M. O., Hoff, S., Jakobs, K., Kubitz, O.: *Directory Services: Basis for Mobile Computing with DECT*, Proc. 5th IEEE International Conference on Universal Personal Communications (ICUPC'96), Cambridge, MA, USA, September 1996, pp. 41–45.
- [BKW93] Bafutto, M., Kühn, P. J., Willmann, G.: *Modelling and Performance Analysis of Common Channel Signalling Networks*, AEÜ (Archiv für Elektronik und Übertragungstechnik), Vol. 47, No. 5, Hirzel-Verlag, 1993, pp. 411–419.
- [BKW94] Bafutto, M., Kühn, P., Willman, G.: *Capacity and Performance Analysis of Signalling networks in Multivendor Environments*, IEEE Journal on Selected Areas in Communications, Vol. 12, No. 3, April 1994, pp. 490–500.

- [BKWZ92] Bafutto, M., Kühn, P. J., Willmann, G., Zepf, J.: *A capacity and performance planning tool for signalling networks based on CCITT signalling system No.7*, Proceedings of the Int. Council for Computer Communication Intelligent Networks Conf. (Ed. P.W. Bayliss), May 1992, pp. 368–379.
- [Bod94] Bodamer, S.: *Mittlere Durchlaufzeiten in gesättigten M/G/1-Prioritätensystemen mit Rückkopplungen*, Diplomarbeit, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, August 1994.
- [Bod96] Bodamer, S.: *Performance Evaluation of Saturated Priority Systems with Feedback*, Proceedings of the 5th Open Workshop on High Speed Networks, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung und Ecole Nationale Supérieure des Télécommunications, März 1996.
- [BPT96] Bhagwat, P., Perkins, C., Tripathi, S.: *Network Layer Mobility: An Architecture and Survey*, IEEE Personal Communications, Vol. 3, No. 3, June 1996, pp. 54–64.
- [BS95] Bafutto, M., Schopp, M.: *Planning the Capacity and Performance of Intelligent Networks based on the ITU-TS IN Capability Set 1*, Proceedings of the Regional ITC-Seminar, Pretoria, South Africa, September 1995, pp. 282–296.
- [BS96] Bafutto, M., Schopp, M.: *Network Performance and Capacity Figures of Intelligent Networks based on the ITU-TS IN Capability Set 1*, Proceedings of the International Workshop on Advanced Intelligent Networks (AIN'96), Passau, Germany, Universität Passau, Fakultät für Mathematik und Informatik, March 1996, pp. 15–29.
- [BS97] Bodamer, S., Schopp, M.: *Erwartungswerte von gesättigten Prioritätensystemen*, Messung, Modellierung und Bewertung von Rechen- und Kommunikationssystemen (MMB'97), Freiberg (Sachsen), Band 1, VDE Verlag GmbH, September 1997, pp. 207–216.
- [BTE88] *CCITT signalling system No.7*, British Telecommunications Engineering, Vol. 7, 1988, pp. 4–71.
- [Buit94] E. Buitenwerf: *Considerations on network interfaces for third generation mobile telecommunication systems*, Proceedings WCN (IEEE/ICCC) 1994, pp. 857–862 WA 3.3.
- [CH94] Colombo, G., Hegeman, H.: *Network architecture and functionalities in UMTS*, Proceedings WCN (IEEE/ICCC) 1994, 1994, pp. 844–851 WA 3.1.
- [Cha81] Chaum, D. L.: *Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms*, Communications of the ACM, Vol. 24, 1981, pp. 84–88.
- [CM65] Cox, D. R., Miller, H. D.: *The theory of stochastic processes*, Chapman and Hall, 1965.
- [Col95] Colombo, G.: *Mobility Models for Mobile System Design and Dimensioning*, Proceedings of the 9th ITC Specialists Seminar: Teletraffic Modelling and Measurement in Broadband and Mobile Communications, Leidschendam, KPN Research, November 1995, pp. 133–146.
- [Com91] Comer, D. E.: *Internetworking with TCP/IP - Volume I, Principles, Protocols and Architecture*, Prentice Hall, 1991.
- [Cou77] Courtois, P. J.: *Decomposability*, Academic Press (ACM monograph series), 1977.
- [CS95] Claus, J., Siegmund, G.: *Das ATM-Handbuch: Grundlagen, Planung, Einsatz*, Hüthig, Heidelberg, 1995.

- [Dat98] Date, C.: *An Introduction to Database Systems*, The Systems Programming Series, Addison-Wesley, April 1998.
- [Dav97] David, K.: *Network Evolution Aspects towards a Third Generation Mobile Systems*, EPMCC '97: Vorträge der 2. EPMCC '97 und 3. ITG-Fachtagung 1997, Bonn, VDE-Verlag GmbH, October 1997.
- [DH97] Darmois, E., Hoshi, M.: *TINA in a nutshell (TINA - A common software architecture for multimedia and information services)*, <http://www.tinac.com/about/nutshell.htm>, February 1997.
- [DK95] Dupuy, F., Kung, R.: *Au-delà du réseau intelligent: le réseau d'information TINA*, L'écho des Recherches, No. 162, Issy-les-Moulineaux, France, pp. 49–60.
- [DNI95] Dupuy, F., Nilsson, G., Inoue, Y.: *The TINA Consortium: Towards Networking Telecommunications Information Services*, IEEE Communications Magazine, Vol. 33, No. 11, November 1995, pp. 78–83.
- [DV92] Duran, J. M., Visser, J.: *International Standards for Intelligent Networks*, IEEE Communications Magazine, Vol. 30, No. 2, February 1992, pp. 34–42.
- [DVS97] Dommety, G., Veeraraghavan, M., Singhal, M.: *Flat Location Management Scheme for PCNs*, Proceedings IEEE 6th International Conference on Universal Personal Communications (ICUPC), San Diego, USA, October 1997, pp. 146–152.
- [E.164] *ITU-T Recommendation E.164/I.331 (08/91) - Numbering plan for the ISDN era*, International Telecommunication Union, Geneva, August 1991.
- [E.212] *ITU-T Recommendation E.212 (1988) - Identification plan for land mobile stations*, International Telecommunication Union, Geneva, 1988.
- [E.214] *ITU-T Recommendation E.214 (1988) - Structure of the land mobile global title for the signalling connection control part (SCCP)*, International Telecommunication Union, Geneva, 1988.
- [EM96] Eckardt, T., Magedanz, T.: *From IN towards TINA-based Personal Communications Support*, Proceedings 5th IEEE International Conference on Universal Personal Communications (ICUPC'96), Cambridge, MA, USA, September 1996, pp. 402–407.
- [EMPZ96] Eckard, T., Magedanz, T., Popescu-Zeletin, R., Schulz, M., Stapf, M.: *Personal Communications Support within the TINA Service Architecture - A new TINA-C Auxiliary Project*, Proceedings TINA'96 Conference, Heidelberg, VDE-Verlag, September 1996, pp. 55–64.
- [ETSI96] *ETSI NA-61301 Document: IN/UMTS Framework v8.2.0*, European Telecommunications Standards Institute, Sophia Antipolis, France, January 1996.
- [EV97] Eberspächer, J., Vögel, H. J.: *GSM: Global System for Mobile Communication*, B.G. Teubner, Stuttgart, Dezember 1997.
- [F.850] *ITU-T Recommendation F.850 (03/93) - Principles of Universal Personal Telecommunication (UPT)*, ITU-T Rec., International Telecommunication Union, Geneva, November 1993.
- [F.851] *ITU-T Recommendation F.851 (02/95) - Universal Personal Telecommunication (UPT) -Service description (service set 1)*, ITU-T Rec., International Telecommunication Union, Geneva, February 1995.

- [FBWC97] Faggion, N., Barry, M., Weber, A., Crowley, C.: *Application of IN Protocols and Framework to Mobility Management in the Rainbow Project*, Proceedings of the ACTS Mobile Communications Summit '97, Aalborg, Danmark, October 1997, pp. 104–109.
- [FGKS97] Faynberg, I., Gabuzda, L. R., Kaplan, M. P., Shah, N. J.: *The Intelligent Network Standards: their application to services*, McGraw-Hill Communications Books, McGraw-Hill, 1997.
- [FHP95] Fasbender, A., Hoff, S., Pietschmann, M.: *Mobility Management in Third Generation Mobile Networks*, Proc. IFIP TC6 Int. Workshop on Personal Wireless Communications (Wireless Local Access) (PWC95): Prague, April 1995, IFIP TC6, April 1995, pp. 33–46.
- [FJKP95] Federrath, H., Jerichow, A., Kesdogan, D., Pfitzmann, A.: *Security in Public Mobile Communication Networks*, Proc. IFIP TC6 Int. Workshop on Personal Wireless Communications (Wireless Local Access) (PWC95): Prague, April 1995, IFIP TC6, April 1995, pp. 105–116.
- [FSY90] Fujioka, M., Sakai, S. I., Yagi, H.: *Hierarchical and Distributed Information Handling for UPT*, IEEE Network Magazine, Vol. 4, No. 6, November 1990, pp. 50–60.
- [G01.02] ETSI: *ETR 099 : European digital cellular telecommunications system (Phase 2), General description of a GSM Public Land Mobile Network (PLMN) GSM Public Land Mobile Network (PLMN) (GSM 01.02)*, European Telecommunications Standards Institute, Sophia Antipolis, France, October 1993.
- [G01.04] ETSI: *ETR 100: European digital cellular telecommunications system (Phase 2), Abbreviations and acronyms (GSM 01.04)*, European Telecommunications Standards Institute, Sophia Antipolis, France, April 1995.
- [G02.79] ETSI: *GTS Digital cellular telecommunications system (Phase 2+), Support of Optimal Routeing (SOR), Service definition (Stage 1) (GSM 02.79 version 5.2.0 Release 1996)*, European Telecommunications Standards Institute, Sophia Antipolis, France, November 1996.
- [G03.79] ETSI: *TS 101 045 V5.4.0: Digital cellular telecommunications system (Phase 2+), Support of Optimal Routeing (SOR), Technical realisation (GSM 03.79 version 5.4.0 Release 1996)*, European Telecommunications Standards Institute, Sophia Antipolis, France, November 1996.
- [G09.02] ETSI: *ETS 300 599 : Digital cellular telecommunications system (Phase 2), Mobile Application Part (MAP) specification (GSM 09.02)*, European Telecommunications Standards Institute, Sophia Antipolis, France, 1995.
- [Gae77] Gaede, K. W.: *Zuverlässigkeit: Mathematische Modelle*, Theorie und Praxis des Operations Research, Carl Hanser Verlag, München, 1977.
- [GMM] ETSI, Rose, G.: *ETSI/PAC16(96)16: Global Multimedia Mobility (GMM) - A Standardization Framework*, European Telecommunications Standards Institute, Sophia Antipolis, France, June 1996.
- [Goy98] Goyal, A.: *Spezifikation und Implementierung eines Protokolls zur Integration von GSM in UPT*, Diplomarbeit, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, Dezember 1998.
- [GRKK93] Garrahan, J. J., Russo, P. A., Kitami, K., Kung, R.: *Intelligent Network Overview*, IEEE Communications Magazine, Vol. 31, No. 3, March 1993, pp. 30–36.

- [GSG97] Garg, V. K., Sneed, E. L., Gooding, W. E.: *Subscriber Data Management in Personal Communications Services Networks*, IEEE Personal Communications, Vol. 4, No. 3, June 1997, pp. 33–39.
- [Gun97] Guntermann, M.: *Universelle Benutzermobilität auf der Basis des Intelligenten Netzes - Entwurf, Bewertung und Implementierung*, Dissertationschrift, RWTH Aachen, Lehrstuhl für Kommunikationsnetze, Dezember 1997.
- [HA95] Ho, J. S. M., Akyildiz, I. F.: *Mobile user location update and paging under delay constraints*, Wireless Networks, Vol. 1, No. 4, ACM-Press, Baltzer Science Publishers, December 1995, pp. 413–425.
- [HA96] Ho, J. S. M., Akyildiz, I. F.: *Local Anchor Scheme for Reducing Signaling Costs in Personal Communications Networks*, IEEE/ACM Transactions on Networking, Vol. 4, No. 5, October 1996, pp. 709–725.
- [HLNS78] Heller, W. D., Lindenberg, H., Nuske, M., Schriever, K. H.: *Stochastische Systeme: Markoffketten, stochastische Prozesse, Warteschlangen*, de Gruyter Lehrbuch, Walter de Gruyter, Berlin, 1978.
- [HMS97] Hellebrandt, M., Mathar, R., Schopp, M.: *Optimum Location Management with Mirror-HLRs*, Interner Bericht, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, Juni 1997.
- [ITG96a] ITG: *Entwurf für die ITG-Empfehlung 5.2-01 "Architekturen und Verfahren der Vermittlungstechnik"* (Stand: Dezember 1996), Begriffswerk des ITG-Fachausschusses 5.2, Informationstechnische Gesellschaft im VDE, Dezember 1996.
- [ITG96b] ITG: *Entwurf für die ITG-Empfehlung 5.2-02 "Systeme der Vermittlungstechnik"* (Stand: Dezember 1996), Begriffswerk des ITG-Fachausschusses 5.2, Informationstechnische Gesellschaft im vde, Dezember 1996.
- [ITG97] ITG: *Entwurf für die ITG-Empfehlung 5.2-03 "Begriffe der Nachrichtenverkehrstheorie"* (Stand: Oktober 1997), Begriffswerk des ITG-Fachausschusses 5.2, Informationstechnische Gesellschaft im VDE, Oktober 1997.
- [ITS98] *International Telecom Statistics 1998 - Internationale Fernmeldestatistik 1998*, Siemens, 1998.
- [JL95] Jain, R., Lin, Y. B.: *An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS*, Wireless Networks, Vol. 1, No. 2, ACM-Press, Baltzer Science Publishers, July 1995, pp. 197–210.
- [JLM94] Jain, R., Lin, Y. B., Mohan, S.: *A Caching Strategy to Reduce Network Impacts of PCS*, IEEE Journal on Selected Areas in Communications, Vol. 12, No. 8, October 1994, pp. 1434–1444.
- [JLM95] Jain, R., Lin, Y. B., Mohan, S.: *A Forwarding Strategy to Reduce Network Impacts of PCS*, Proceedings of IEEE Infocom '95: The Conference on Computer Communications, 14th Annual Joint Conference of the IEEE Computer and Communications Societies, Boston, MA, USA, IEEE Computer Society Press, April 1995, pp. 481–489.
- [Kan94] Kandulski, A.: *Untersuchung des Signalisiermeldungsflusses im Intelligenten Netz*, Diplomarbeit, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, September 1994.

- [KCU95] Kim, B. C., Choi, J. S., Un, C. K.: *A New Distributed Location Management Algorithm for Broadband Personal Communication Networks*, IEEE Transactions on Vehicular Technology, Vol. 44, No. 3, August 1995, pp. 516–524.
- [Kes96] Kesper, K.: *Verteilte Verzeichnissysteme und replizierte Datenbanken*, Dissertationsschrift, Philipps-Universität Marburg, Fakultät Mathematik, 1996.
- [KFJP] Kesdogan, D., Federrath, H., Jerichow, A., Pfitzmann, A.: *Location Management Strategies increasing Privacy in Mobile Communication Systems*, Proceedings of the IFIP SEC '96, 12th International Information Security Conference, May 1996, Chapman & Hall, pp. 39–48.
- [KK98] Keck, D. O., Kühn, P. J.: *The Feature and Service Interaction Problem in Telecommunications Systems: A Survey*, IEEE Transactions on Software Engineering, Vol. 24, No. 10, October 1998, pp. 779–796.
- [KKR97] Kotsis, G., Krithivasan, K., Raghavan, S. V.: *A Workload Characterization Methodology for WWW Applications*, Proceedings of the International Conference on the Performance and Management of Complex Communication Networks (PMCCN'97), Tsukuba, Japan, University of Tsukuba, Kyoto University, November 1997, pp. 145–159.
- [Kle97] Kleier, S.: *Neue Konzepte zur Unterstützung von Mobilität in Telekommunikationsnetzen*, Dissertationsschrift, RWTH Aachen, Lehrstuhl für Kommunikationsnetze, Dezember 1997.
- [KM94] Kung, R., Maitre, X.: *L'architecture du réseau intelligent*, L'écho des Recherches, No. 157, Issy-les-Moulineaux, France, 1994, pp. 5–14.
- [KPS94] Kühn, P. J., Pack, C. D., Skoog, R. A.: *Common Channel Signaling Networks: Past, Present, Future*, IEEE Journal on Selected Areas in Communications, Vol. 12, No. 3, April 1994.
- [KS76] Kemeny, J. G., Snell, J. L.: *Finite Markov Chains*, Undergraduate Texts in Mathematics, Springer, Berlin, 1976.
- [KS96] Kühn, P., Schopp, M.: *Signalling Networks for ISDN, IN and Mobile Networks - Modelling, Analysis and Overload Control*, Proceedings of the 10th ITC Specialists Seminar on Control in Communications, Lund, Sweden, 1996, Lund University, September 1996, pp. 35–49.
- [Kü76] Kühn, P. J.: *Analysis of complex queueing networks by decomposition*, Proceedings of the 8th International Teletraffic Congress, 1976, pp. 1–8.
- [Kü77] Kühn, P. J.: *Approximate analysis of general queueing networks by decomposition*, IEEE Transactions on Communications, No. 25, 1977, pp. 113–126.
- [Kü96] Kühn, P. J.: *Unterlagen zur Lehrveranstaltung "Nachrichtenverkehrstheorie"*, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, 1996.
- [Kwi96] Kwiatkowski, M.: *Calculating Rates of Services in UPT Networks*, Proc. 5th IEEE International Conference on Universal Personal Communications (ICUPC'96), Cambridge, MA, USA, September 1996, pp. 513–517.
- [Law95] Lawniczak, D. R.: *Modellierung und Bewertung der Datenverwaltungskonzepte in UMTS*, Dissertationsschrift, RWTH Aachen, Lehrstuhl für Kommunikationsnetze, 1995.
- [Leu97] Leung, K. K.: *An Update Algorithm for Replicated Signaling Databases in Wireless and Advanced Intelligent Networks*, IEEE Transactions on Computers, Vol. 46, No. 3, March 1997, pp. 362–367.

- [LH96] Lin, Y. B., Hwang, S. Y.: *Comparing the PCS Location Tracking Strategies*, IEEE Transactions on Vehicular Technology, Vol. 45, No. 1, February 1996, pp. 114–121.
- [Lin97] Lin, Y. B.: *Reducing Location Update Cost in a PCS Network*, IEEE/ACM Transactions on Networking, Vol. 5, No. 1, February 1997, pp. 25–33.
- [Lit63] Little, J. D. C.: *A proof for the queueing formula $L = \lambda * W$* , Operations Research, 1963, pp. 383–387.
- [LJCW96] Lam, D., Jannink, J., Cox, D. C., Widom, J.: *Modeling Location Management in Personal Communication Services*, Proc. 5th IEEE International Conference on Universal Personal Communications (ICUPC'96), Cambridge, MA, USA, September 1996.
- [LL83] Laning, L., Leonard, M.: *File Allocation in a Distributed Computer Communication Network*, IEEE Transactions on Computers, Vol. 32, No. 7, 1983, pp. 232–244.
- [LL97a] Leung, K. K., Levy, Y.: *Global Mobility Management by Replicated Databases in Personal Communication Networks*, IEEE Journal on Selected Areas in Communications, Vol. 15, No. 8, October 1997, pp. 1582–1596.
- [LL97b] Leung, K. K., Levy, Y.: *Data Replication Schemes for Global Networks*, Proceedings of the 15th International Teletraffic Congress (ITC 15): Teletraffic Contributions for the Information Age, Washington, DC, USA, June 1997, Elsevier, pp. 211–222.
- [LMMS97] Licciardi, C. A., Minerva, R., Moiso, C., Spinelli, G.: *Would you use TINA in your IN based Network? Some feasible scenarios*, CSELT Technical reports, Vol. 25, No. 1, February 1997, pp. 9–26.
- [LS96] Lautenschlager, W., Stahl, U.: *Routing Service for the Implementation of Number Portability*, Proceedings of 4th International Conference on Intelligence in Networks (ICIN'96), November 1996, pp. 19–24.
- [LTU95] *Zeichengabesysteme: eine neue Generation für ISDN und intelligente Netze*, LTU-Vertriebsgesellschaft mbH, Bremen, 1995.
- [LY96] Lindholm, T., Yellin, F.: *The Java Virtual Machine Specification*, Addison Wesley Longman, 1996.
- [M.3000] *ITU-T Recommendation M.3000 (10/94) - Overview of TMN Recommendations*, International Telecommunication Union, Geneva, October 1994.
- [Man91] Manterfield, R. J.: *Common-Channel Signalling.*, Peter Peregrinus Ltd., London, 1991.
- [MH98] Mathar, R., Hellebrandt, M.: *Optimum Mirror-HLR locations to reduce signalling load in cellular radio networks*, Multiaccess, Mobility and Teletraffic: Advances in Wireless Networks, Melbourne, Australia, 01.12.1997 - 12.12.1997, Klumer Academic Publishers, March 1998, pp. 139–250.
- [MJ94] Mohan, S., Jain, R.: *Two User Location Strategies for Personal Communications Services*, IEEE Personal Communications, Vol. 1, No. 1, 1994, pp. 42–50.
- [MM96] Minetti, R., Mulder, H.: *Overview of TINA: Tutorial B of the Intelligent Network '96*, Sunday, 21 April 1996, Carlton Crest Hotel, Melbourne, Australia, April 1996.
- [MN96] Morris, C., Nelson, J.: *UMTS Architectures for the Support of UPT in UMTS*, Proceedings of ACTS Mobile Telecommunications Summit: Volume 2, Granada, Spain, November 1996, pp. 811–820.
- [MON94] MONET: *Implementation Aspects of the UMTS Database*, RACE Deliverable R2066/PTTNL/MF/DS/P/061/b1, May 1994.

- [MON95a] MONET: *Baseline document on logical and functional models*, RACE Deliverable R2066/PTTNL/MF/DS/P/099/a2, December 1995.
- [MON95b] MONET: *Distributed Database for UMTS and Integration with UPT*, RACE Deliverable R2066/SEL/UNA3/DS/P/108/b1, December 1995.
- [MP92] Mouly, M., Pautet, M. B.: *The GSM System for Mobile Communications*, published by the authors, 1992.
- [MPZ96] Magedanz, T., Popescu-Zeletin, R.: *Intelligent Networks: Basic Technology, Standards and Evolution*, International Thomson Computer Press, London, 1996.
- [OMG97] OMG: *The Common Object Request Broker: Architecture and Specification (CORBA/IIOP 2.1 Specification)*, August 1997.
- [ÖV91] Özsu, M. T., Valduriez, P.: *Principles of distributed database systems*, Prentice-Hall, London, 1991.
- [Pat90] Paterok, M.: *Warteschlangensysteme mit Rückkopplung und Prioritäten*, Dissertationsschrift, Friedrich-Alexander-Universität Erlangen-Nürnberg, IMMD (Informatik), 1990.
- [Per96] Perkins, C. E.: *IP Mobility Support (Mobile IP) RFC 2002*, Internet Engineering Task Force, October 1996.
- [Per97] Perkins, C. E.: *Mobile IP*, IEEE Communications Magazine, Vol. 35, No. 5, May 1997, pp. 84–99.
- [PF89] Paterok, M., Fischer, O.: *Feedback queues with preemption distance priorities*, ACM Sigmetrics Performance Evaluation Review, Vol. 17, No. 1, 1989, pp. 136–145.
- [Pfi93] Pfitzmann, A.: *Technischer Datenschutz in öffentlichen Funknetzen*, Datenschutz und Datensicherung, DuD, No. 8, 1993, pp. 451–463.
- [PJ97] Perkins, C., Johnson, D. B.: *Route Optimization in Mobile IP*, Internet Draft, draft-ietf-mobileip-optim-07.txt, IETF Mobile IP Working Group, November 1997.
- [Pla94] Plassmann, D.: *Location management for MBS*, Proceedings RACE Mobile Telecommunications Workshop, Amsterdam, May 1994, pp. 418–423.
- [Pol96] Polat, A.: *Integration eines Werkzeugs zur Leistungsuntersuchung von Intelligenten Netzen*, Studienarbeit, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, Mai 1996.
- [PPW89] Pfitzmann, A., Pfitzmann, B., Waidner, M.: *Telefon-MIXe: Schutz der Vermittlungsdaten für zwei 64-kbit/s-Duplexkanäle über den (2*64+16)-kbit/s-Teilnehmeranschluß*, Datenschutz und Datensicherung, DuD, No. 12, 1989, pp. 605–622.
- [PPW91] Pfitzmann, A., Pfitzmann, B., Waidner, M.: *ISDN-MIXes: Untraceable Communication with Very Small Bandwidth Overhead*, Kommunikation in verteilten Systemen KiVS'91 Grundlagen, Anwendungen, Betrieb, GI/ITG-Fachtagung, Springer-Verlag, February 1991, pp. 451–463.
- [Q.700] ITU-T Recommendation Q.700 (03/93) - *Introduction to CCITT Signalling System No. 7*, International Telecommunication Union, Geneva, March 1993.
- [Q.701] ITU-T Recommendation Q.701 (03/93) - *Functional description of the message transfer part (MTP) of Signalling System No. 7*, International Telecommunication Union, Geneva, March 1993.
- [Q.711] ITU-T Recommendation Q.711 (07/96) - *Signalling System No. 7 - Functional description of the Signalling Connection Control Part*, International Telecommunication Union, Geneva, July 1996.

- [Q.761] *ITU-T Recommendation Q.761 (03/93) - Functional description of the ISDN user part of Signalling System No. 7*, International Telecommunication Union, Geneva, March 1993.
- [Q.771] *ITU-T Recommendation Q.771 (06/97) - Signalling System No. 7 - Functional description of Transaction Capabilities*, International Telecommunication Union, Geneva, June 1997.
- [Q.1200] *ITU-T Recommendation Q.1200 (03/93) - Q-Series Intelligent Network Recommendation structure*, International Telecommunication Union, Geneva, March 1993.
- [Q.1201] *ITU-T Recommendation I.312/Q.1201 (10/92) - Principles of intelligent network architecture*, International Telecommunication Union, Geneva, October 1992.
- [Q.1211] *ITU-T Recommendation Q.1211 (03/93) - Introduction to intelligent network capability set 1*, ITU-T Rec., International Telecommunication Union, Geneva, March 1993.
- [Q.1214] *ITU-T Recommendation Q.1214 (10/95) - Distributed functional plane for intelligent network CS-1*, ITU-T Rec., International Telecommunication Union, Geneva, October 1995.
- [Q.1220] *Draft new ITU-T Recommendation Q.1220 proposed for approval under resolution 1 (Geneva, 1996) [COM11-R39] Report of the meeting held in Geneva from 13-31 January 1997 - Part II*, ITU-T Report, International Telecommunication Union, Geneva, January 1997.
- [Q.1221] *Draft new ITU-T Recommendation Q.1221 proposed for approval under resolution 1 (Geneva, 1996) [COM11-R40] Report of the meeting held in Geneva from 13-31 January 1997 - Part II*, ITU-T Report, International Telecommunication Union, Geneva, January 1997.
- [Q.1224] *Draft new ITU-T Recommendation Q.1224 proposed for approval under resolution 1 (Geneva, 1996) [COM11-R43,R44,R45,R46] to Report of the meeting held in Geneva from 13-31 January 1997 - Part II*, ITU-T Report, International Telecommunication Union, Geneva, January 1997.
- [Q.2761] *ITU-T Recommendation Q.2761 (02/95) - Broadband integrated services digital network (B-ISDN) - Functional description of the B-ISDN user part (B-ISUP) of signalling system No. 7*, International Telecommunication Union, Geneva, February 1995.
- [RH95] Reichardt, U., Heilig, B.: *Service Creation im Intelligenten Netz (IN)- Der Schlüssel zum Erfolg*, Taschenbuch der Telekom Praxis 1995, Schiele & Schön, 1995.
- [RT96] Raatikainen, K., Tuoriniemi, A.: *Mobility in the Intelligent Network Architecture*, Proceedings of 4th International Conference on Intelligence in Networks (ICIN'96), November 1996, pp. 289–294.
- [SB95] Schopp, M., Bafutto, M.: *Untersuchung des Signalisiermeldungsflusses bei IN-basierter Mobilitätsverwaltung*, ITG-Fachbericht 135: Mobile Kommunikation: Vorträge der ITG-Fachtagung 1995 in Neu-Ulm, VDE-Verlag, September 1995, pp. 215–222.
- [SB97] Schmidt, S., Breuer, W.: *Zukünftige Numerierung im Telefon-/ISDN-Netz in der Bundesrepublik Deutschland*, Taschenbuch der Telekom Praxis 1997: 34. Jg., Schiele & Schön, 1997, pp. 53–76.
- [Sch96] Schopp, M.: *Network Design for IN-based Location Management*, International Teletraffic Congress Mini-Seminar on Engineering and Congestion Control in Intelligent Networks: Workshop Record, Carlton Crest Hotel, Melbourne, Australia, April 1996, paper 7.

- [Sch97a] Schopp, M.: *User Modelling and Performance Evaluation of Distributed Location Management for Personal Communications Services*, Proceedings of the 15th International Teletraffic Congress (ITC 15): Teletraffic Contributions for the Information Age, Washington, DC, USA, June 1997, Elsevier, pp. 23–34.
- [Sch97b] Schopp, M.: *UPT-GSM Interworking: Support of Multiple Registrations*, Interner Bericht, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, Stuttgart, April 1997.
- [SG10Q9] ITU-TS Study Group 10 (MSC): *Question 9: Correction list for Z.120 (MSC'96)*, April 1997.
- [SM98] Schopp, M., Mayer, S.: *Location Management in a Multi-Network-Operator Environment*, Multimedia and nomadic communications: Proceedings of the 11th ITC Specialist Seminar, Yokohama, Japan, October 1998, pp. 285–292.
- [Tab95] Tabbane, S.: *An Alternative Strategy for Location Tracking*, IEEE Journal on Selected Areas in Communications, Vol. 13, No. 5, June 1995, pp. 880–892.
- [TG97] Tran-Gia, P.: *Analytische Leistungsbewertung verteilter Systeme: eine Einführung*, Springer, Berlin, Oktober 1997.
- [Thö94] Thörner, J.: *Intelligent Networks*, Artech House, 1994.
- [TINA-SA5] Abarca, C., Farley, P., Forslww, J., Garcia, J. C., Hamada, T., Hansen, P. F., Hogg, S., Kamata, H., Kristiansen, L., Licciardi, C. A., Mulder, H., Utsunomiya, E., Yates, M.: *TINA-C Deliverable: Service Architecture Version 5.0*, TINA-C, June 1997.
- [TN95] Tabbane, S., Nevoux, R.: *An intelligent location tracking method for personal and terminal FPLMTS/UMTS communications*, Proceedings of the XV International Switching Symposium (ISS) 1995, Volume 1, Berlin, VDE-Verlag, April 1995, paper A2.1.
- [U30.01] ETSI: *UMTS 30.01 V3.3.0 (1998-06): UMTS Baseline document, Positions on UMTS agreed by SMG*, European Telecommunications Standards Institute, Sophia Antipolis, France, June 1998.
- [VD96a] Veeraraghavan, M., Dommety, G.: *ATM Forum 96-1500: Location Management in Wireless ATM Networks*, ATM Forum, October 1996.
- [VD96b] Veeraraghavan, M., Dommety, G.: *ATM Forum 96-1701: Location management update*, ATM Forum, December 1996.
- [VH98] Venieris, I. S., Hussmann, H.: *Intelligent Broadband Networks*, John Wiley & Sons, 1998.
- [vSHT96] van Steen, M., Hauck, F. J., Tanenbaum, A. S.: *A Model for Worldwide Tracking of Distributed Objects*, Proceedings TINA'96 Conference, Heidelberg, VDE-Verlag, September 1996, pp. 203–212.
- [Wan93] Wang, J. Z.: *A Fully Distributed Location Registration Strategy for Universal Personal Communication Systems*, IEEE Journal on Selected Areas in Communications, Vol. 11, No. 6, August 1993, pp. 850–860.
- [Whi83] Whitt, W.: *Approximations for networks of queues*, Proceedings of the 10th International Teletraffic Congress, 1983, pp. 1–5.
- [Wil88] Willmann, G.: *Modelling and performance evaluation of multi-layered signalling networks based on the CCITT No.7 specification*, Proceedings of the 12th International Teletraffic Congress (ITC), June 1988, pp. 930–940.

- [Wil92] Wilson, D. R.: *Signaling system No.7, IS-41 and cellular telephony networking*, Proceedings of the IEEE, Vol. 80, 1992, pp. 644–652.
- [Wit96] Witte, E.: *Das Telekommunikationsgesetz 1996 - Eine Herausforderung für Markt und Ordnungspolitik*, Münchener Kreis, R. v. Decker's Verlag, Heidelberg, 1996.
- [WK90] Willmann, G., Kühn, P. J.: *Performance Modeling of Signaling System No. 7*, IEEE Communications Magazine, Vol. 28, 1990, pp. 44–56.
- [Wol82] Wolff, R. W.: *Poisson arrivals see time averages.*, Operations Research, Vol. 30, 1982, pp. 223–231.
- [X.217] *ITU-T Recommendation X.217 (04/95) - Information technology - Open systems interconnection - Service definition for the association control service element*, International Telecommunication Union, Geneva, April 1995.
- [X.219] *ITU-T Recommendation X.219 (1988) - Remote operations: Model, notation and service definition*, International Telecommunication Union, Geneva, 1988.
- [X.500] *ITU-T Recommendation X.500 (11/93) - Information technology - Open Systems Interconnection - The directory: Overview of concepts, models, and services*, ITU-T Rec., International Telecommunication Union, Geneva, November 1993.
- [X.518] *ITU-T Recommendation X.518 (11/93) - Information technology - Open Systems Interconnection - The directory: Procedures for distributed operation*, ITU-T Rec., International Telecommunication Union, Geneva, November 1993.
- [X.525] *ITU-T Recommendation X.525 (11/93) - Information technology - Open Systems Interconnection - The directory: Replication*, ITU-T Rec., International Telecommunication Union, Geneva, November 1993.
- [X.650] *ITU-T Recommendation X.650 (10/96) - Information technology - Open Systems Interconnection - Basic Reference Model: Naming and addressing*, ITU-T Rec., International Telecommunication Union, Geneva, October 1996.
- [X.901] *ITU-T Recommendation X.901 (1997) - information technology - basic reference model of open distributed processing - overview*, ITU-T Rec., International Telecommunication Union, Geneva, 1997.
- [X.902] *ITU-T Recommendation X.902 (11/95) - Information technology - Open distributed processing - Reference model: Foundations*, ITU-T Rec., International Telecommunication Union, Geneva, November 1995.
- [X.903] *ITU-T Recommendation X.903 (11/95) - Information technology - Open distributed processing - Reference model: Architecture*, ITU-T Rec., International Telecommunication Union, Geneva, November 1995.
- [Z.120] *ITU-T Recommendation Z.120 (10/96) - Message sequence chart (MSC)*, ITU-T Rec., International Telecommunication Union, Geneva, October 1996.
- [Zepf95] Zepf, J.: *Modellierung und Bewertung von Überlastabwehrmechanismen in Signalsystemen*, Dissertationsschrift, 61. Bericht über verkehrstheoretische Arbeiten, Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, 1995.

Anhang A

A.1 Erweitertes Beispiel zur Parametrisierung des stationären Zustandsmodells

Das Modellierungsbeispiel von Abbildung 5-3a kann auch anders als in Abbildung 5-3b modelliert werden. Anstatt Personen in einer Art von Zufallsspaziergang (*random walk*) im Gang zu bewegen, kann man auch gezielte Übergänge zwischen Büros modellieren. Der Zustandsübergangsgraph dieser Alternative ist in Abbildung A-1 dargestellt.

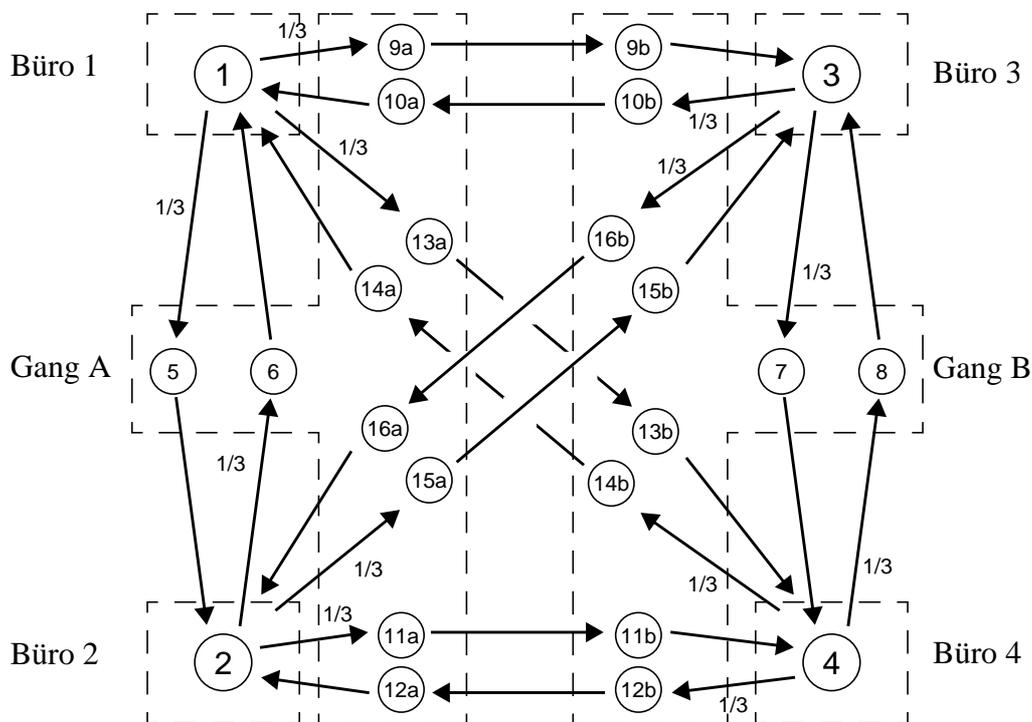


Abbildung A-1: Modellierungsalternative mit gezielten Übergängen zwischen den Büros.

Die Zustände 5 bis 16 repräsentieren alle möglichen direkten Übergänge zwischen den Büros 1 bis 4. Die Übergänge erfolgen ohne daß auf dem Weg zwischen den Büros umgekehrt werden kann. Bei Übergängen, die durch zwei Gangsegmente führen, werden die zugehörigen Zustände in Unterzustände mit Index a und b aufgespalten. Die Übergänge zwischen Unterzuständen sind zur Abbildung auf Übergänge zwischen Aufenthaltsbereichen notwendig. Zur Parametrisierung des Zustandsübergangsgraphen ist diese Aufspaltung nicht notwendig. Ent-

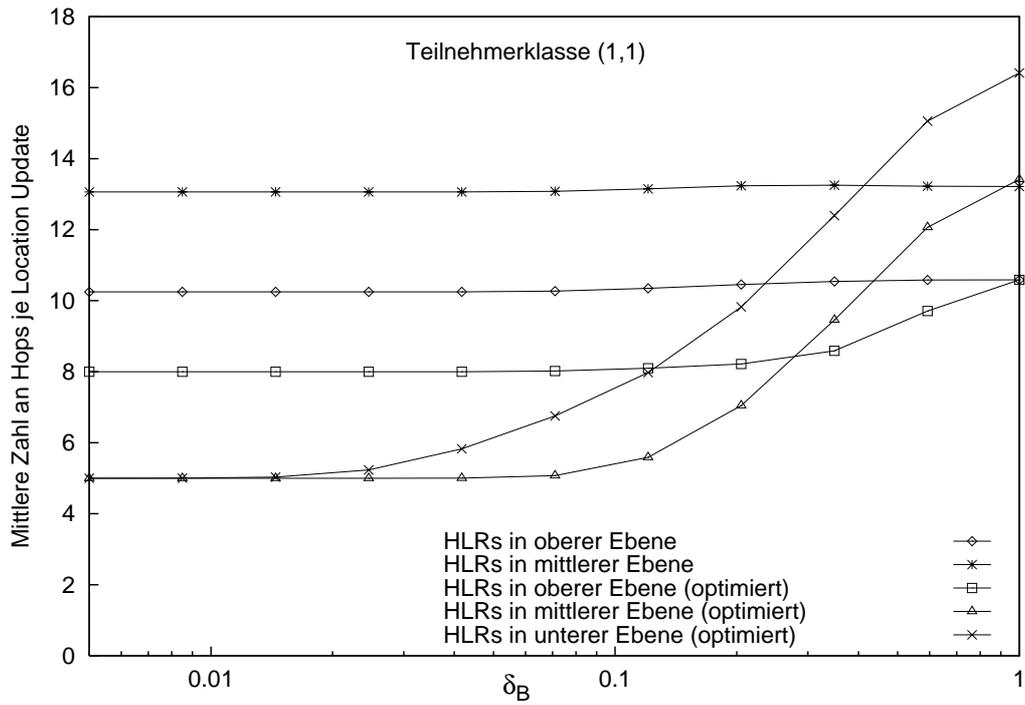


Abbildung A-2: Gleiche Darstellung wie in Abbildung 6-2 für die Teilnehmerklasse mit Heimatknoten (1,1).

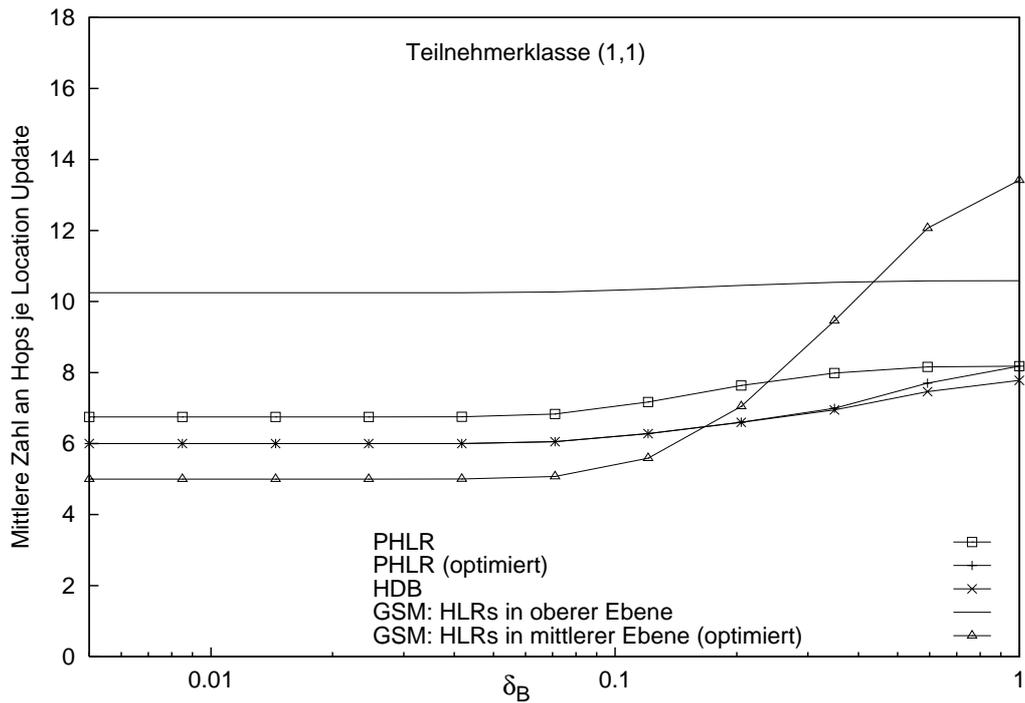


Abbildung A-3: Gleiche Darstellung wie in Abbildung 6-6 für die Teilnehmerklasse mit Heimatknoten (1,1).

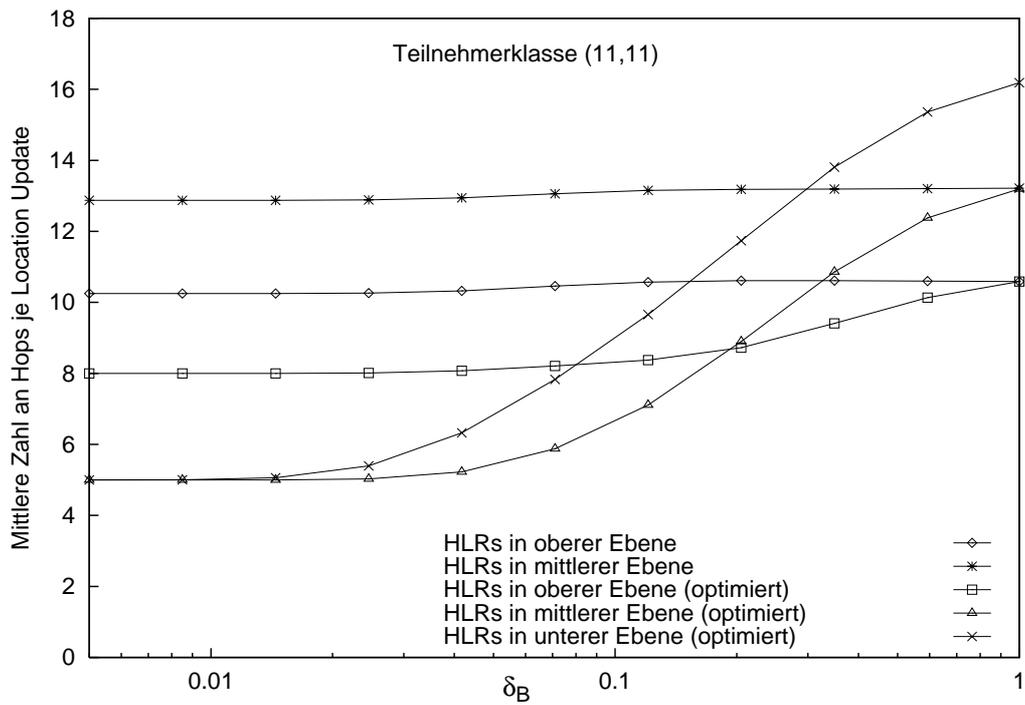


Abbildung A-4: Gleiche Darstellung wie in Abbildung 6-2 für die Teilnehmerklasse mit Heimatknoten (11,11).

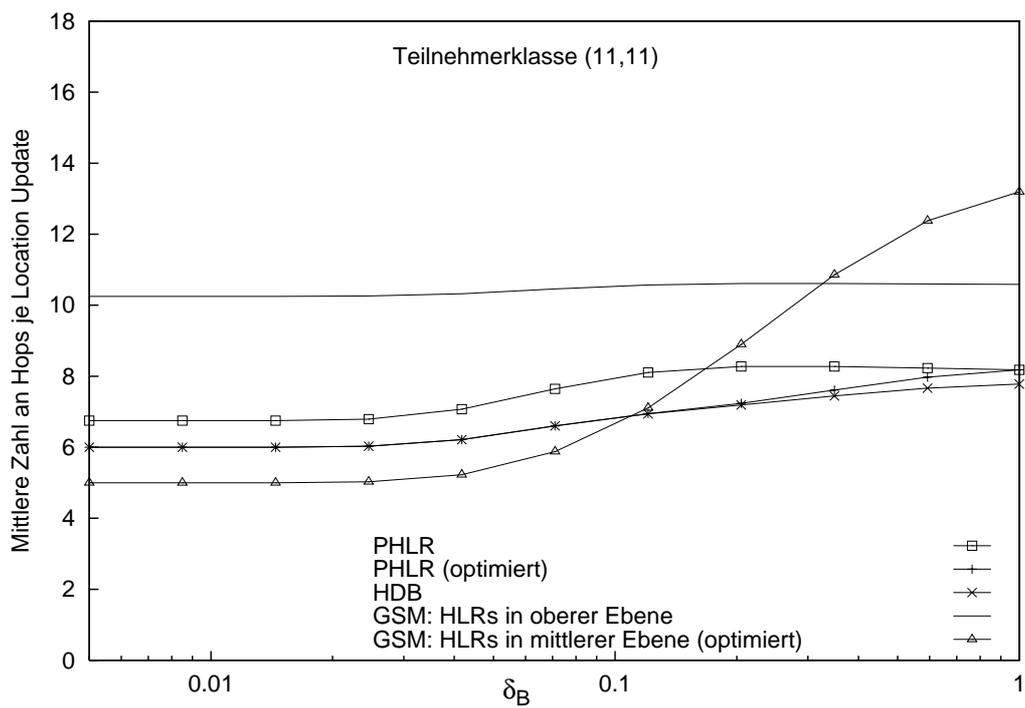


Abbildung A-5: Gleiche Darstellung wie in Abbildung 6-6 für die Teilnehmerklasse mit Heimatknoten (11,11).

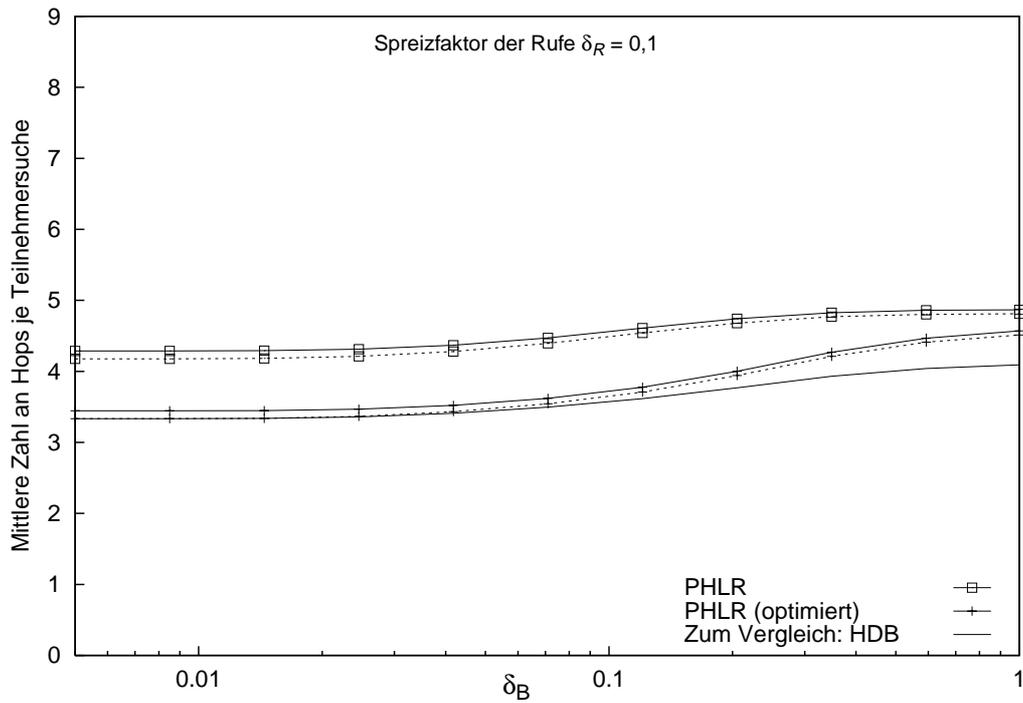


Abbildung A-6: Vergleich zwischen dem *Proxy-HLR*-Verfahren mit „VLR-Lookup“ (gestrichelt) und demselben Verfahren ohne „VLR-Lookup“ (durchgezogen) bezüglich der mittleren Zahl von „Hops“ je Teilnehmersuche mit $\delta_R=0,1$.

Platz absichtlich leer gelassen

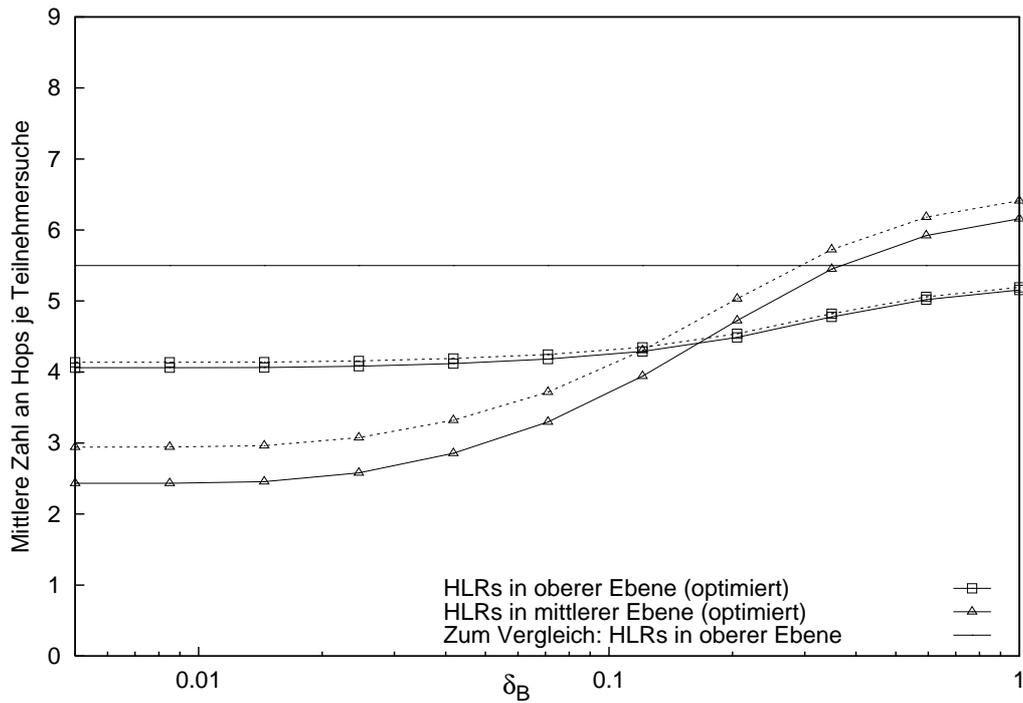


Abbildung A-7: Mittleren Zahl an „Hops“ je Teilnehmersuche bei verschiedenen GSM-Konfigurationen für $\delta_R=0,05$ (durchgezogen) und $\delta_R=0,1$ (gestrichelt).

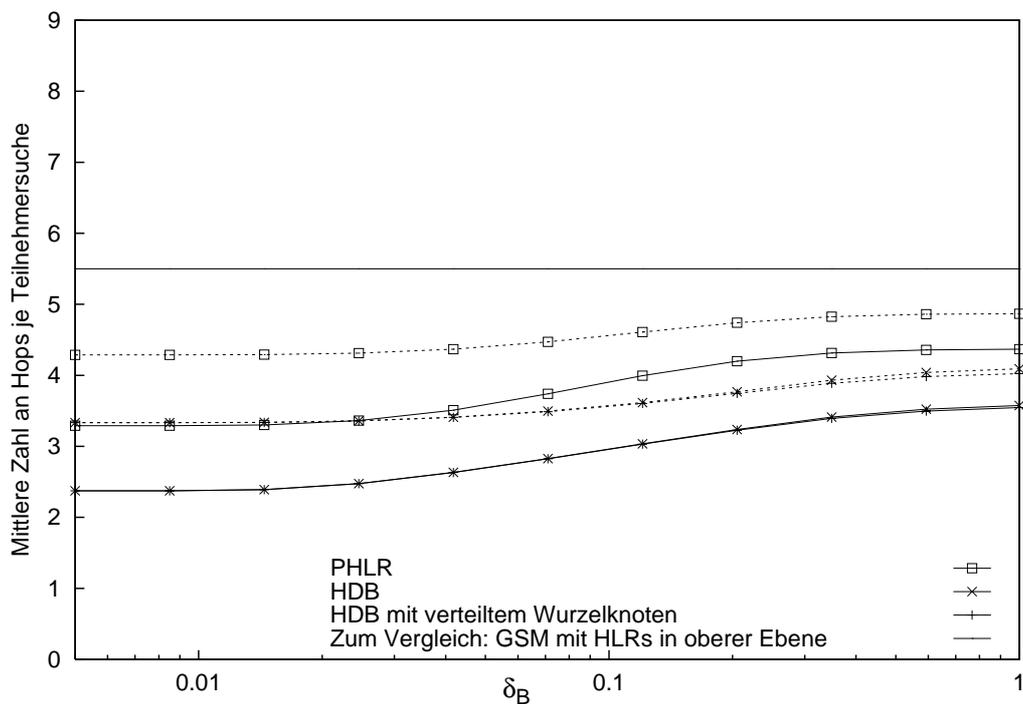


Abbildung A-8: Mittlere Zahl an „Hops“ je Teilnehmersuche bei verschiedenen Mobilitätsverwaltungsverfahren für $\delta_R=0,05$ (durchgezogen) und $\delta_R=0,1$ (gestrichelt).

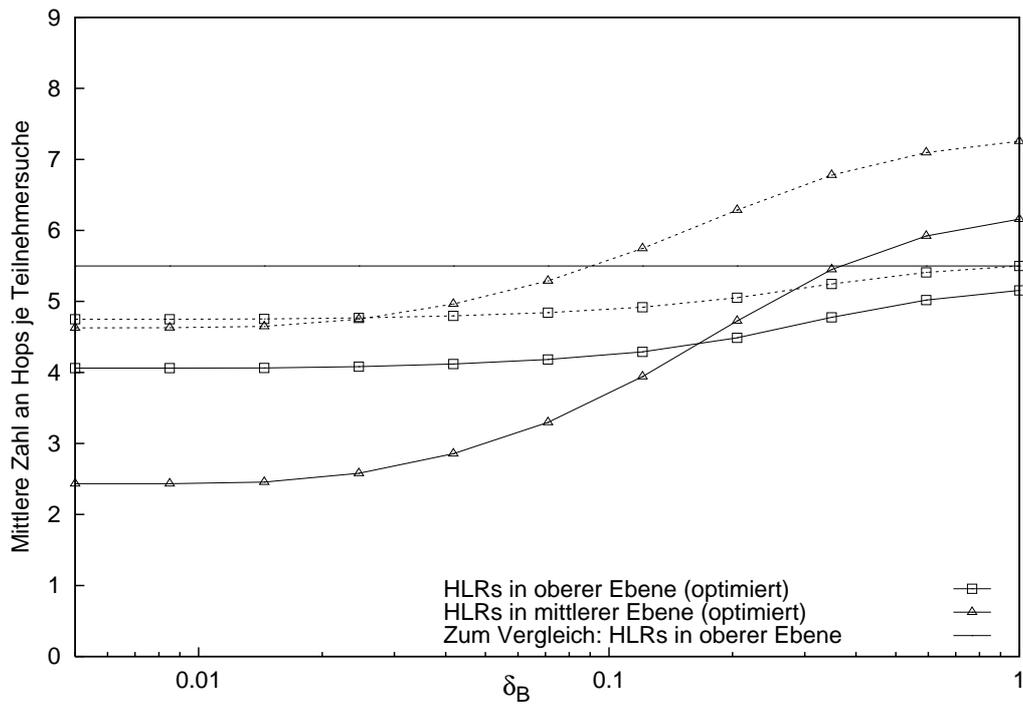


Abbildung A-9: Mittlere Zahl an „Hops“ je Teilnehmersuche bei verschiedenen GSM-Konfigurationen für $\delta_R=0,05$ (durchgezogen) und $\delta_R=1,0$ (gestrichelt).

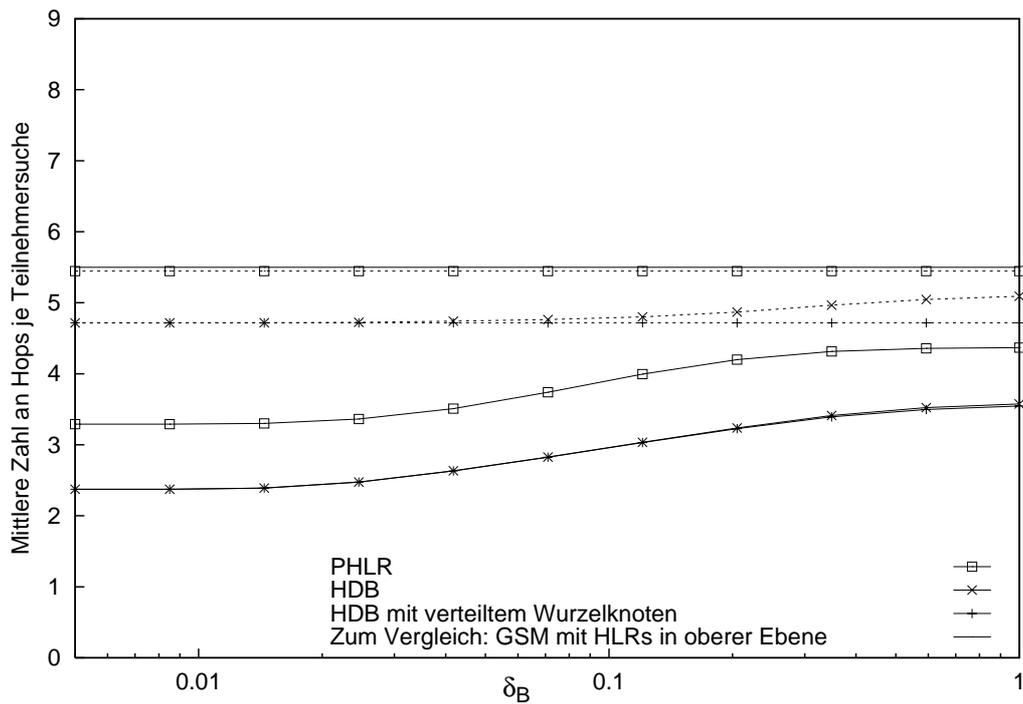


Abbildung A-10: Mittlere Zahl an „Hops“ je Teilnehmersuche bei verschiedenen Mobilitätsverwaltungsverfahren für $\delta_R=0,05$ (durchgezogen) und $\delta_R=1,0$ (gestrichelt).

Anhang B

B.1 Das SSF/CCF Modell im IN CS-1

Abbildung B-1 stellt das gemeinsame Modell von SSF und CCF dar. Via die SSF erhält die Dienststeuerung Zugriff auf die Rufsteuerung eines Halbrufes. Die Sicht der SCF auf den Halbruf ergibt sich im wesentlichen aus dem *Basic Call State Model* (BCSM).

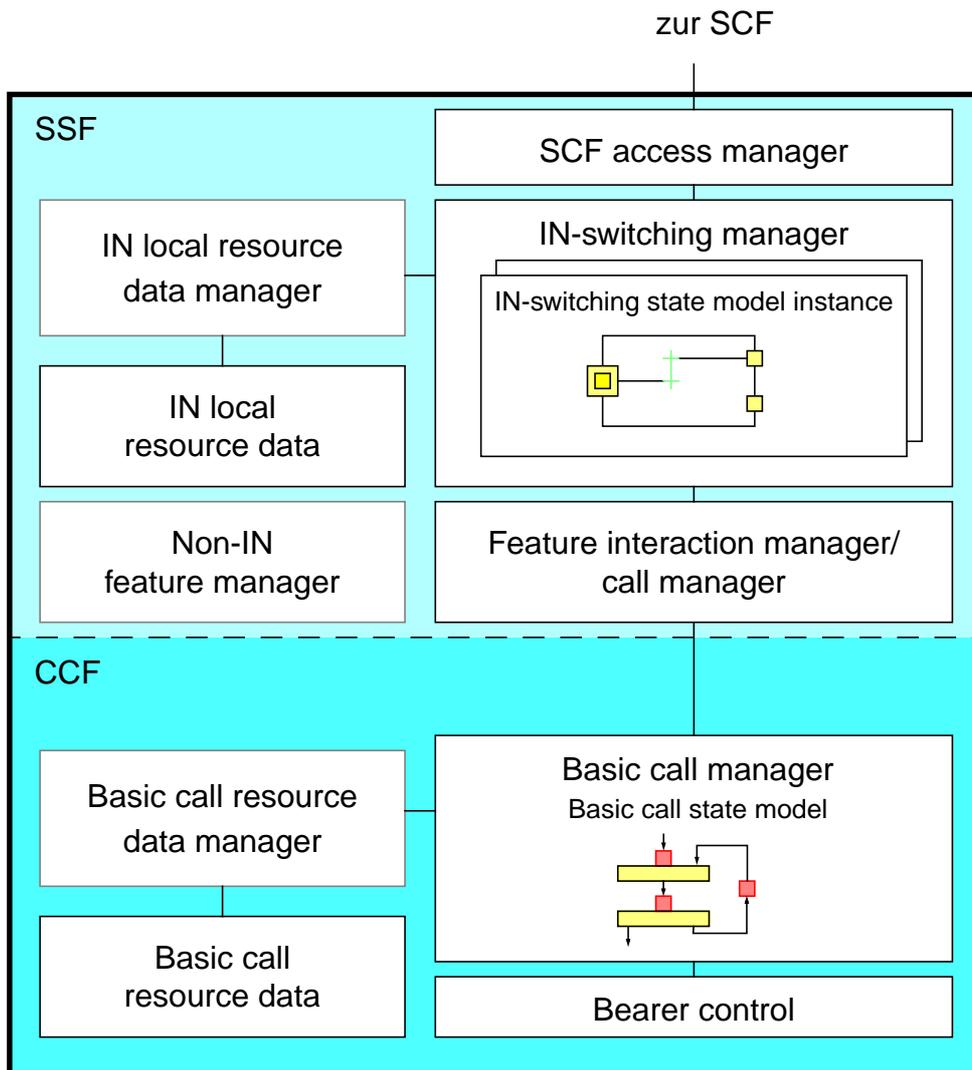


Abbildung B-1: Das SSF/CCF Modell im IN CS-1 für einen Halbruf

B.2 Die zugehörigen *Basic Call State Models*

Die Abbildungen B-2 und B-3 stellen die BCSMs des IN CS-1 mit ihren *Points in Call* und ihren *Detection Points*, an denen Dienstlogik eingreifen kann, dar.

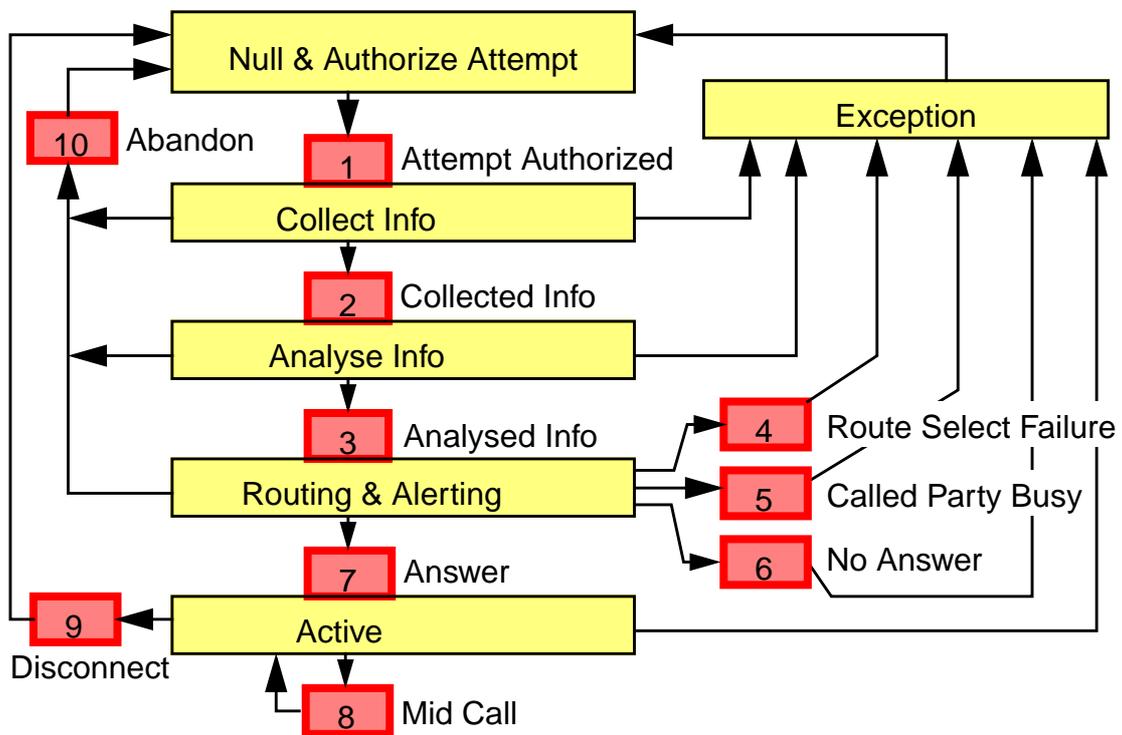


Abbildung B-2: Das *Basic Call State Model* für kommende (Halb-) Rufe

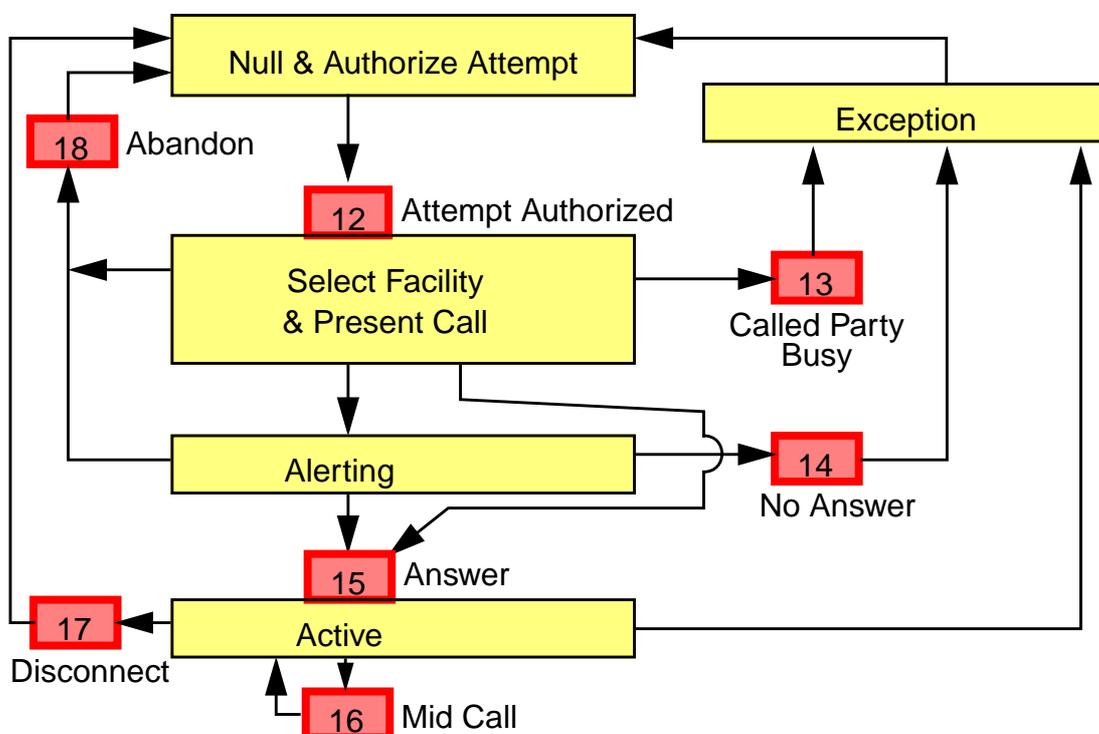


Abbildung B-3: Das *Basic Call State Model* für gehende (Halb-) Rufe